Arbitrary Precision Error Analysis for computing $\zeta(s)$ with the Cohen-Olivier algorithm: Complete description of the real case and preliminary report on the general case

Y.-F.S. Pétermann, Jean-Luc Rémy

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Arbitrary Precision Error Analysis for computing $\zeta(s)$ with the Cohen-Olivier algorithm: Complete description of the real case and preliminary report on the general case

Y.-F.S. Pétermann — Jean-Luc Rémy

N° 5852

February 2006

Thème NUM

ISSN 0249-6399 ISRN INRIA/RR--5852--FR+ENG

INRIA
LORRAINE

Arbitrary Precision Error Analysis for computing $\zeta(s)$ with the Cohen-Olivier algorithm: Complete description of the real case and preliminary report on the general case

Y.-F.S. Pétermann* , Jean-Luc Rémy

Thème NUM — Systèmes numériques
Projet SPACES

Rapport de recherche n° 5852 — February 2006 — 31 pages

Abstract: **1.** Let s be a real number. We prove that, if $s \geq 1/2$, $s \neq 1$ and s can be written with D_s bits in base 2, then in order to compute $\zeta(s)$ in any relative precision $P \geq 11$, that is, in order to compute a $P-$bit number $\zeta_P(s)$ such that $|\zeta_P(s) - \zeta(s)|$ is certified to be smaller than the number $ulp(\zeta_P(s))$ represented by a "1" at the $P-$th (and last) significant bit-place of $|\zeta_P(s)|$, it is sufficient to perform all the computations (i.e. additions, subtractions, multiplications, divisions, and computation of k^s for integers $k \geq 2$) with an internal precision

$$D = \max\left(D_s, P + \max\left(14, \left\lceil \frac{3\log P}{2\log 2} + 2.71 \right\rceil\right)\right)$$

(all this contributing an error less than $ulp(\zeta_P(s)/2)$, and then to round to the nearest $P-$bit number. For instance if the wanted precision is $P = 1000$ (and if s has no more than 1018 significant bits), then an internal precision $D = 1018$ is sufficient.

2. Let $s = \sigma + it$ be a complex non real number. Assume $\sigma \geq 1/2$ and $t > 0$. First we address the problem of exploiting an error relative to modulus in order to estimate the relative errors of each of the real and imaginary parts of the computed $\zeta(s)^*$. Determining regions of the complex plane where these parts cannot vanish could help. Then we establish an easily computable upper bound for a crucial quantity in the error analysis (for the error relative to modulus), subject to the truth of an open conjecture of Brent on the size of the error committed while computing the Bernoulli numbers; we note that the upper bound one can obtain without this conjecture can become so large that even for certain "reasonable" value of s it is of no practical use.

* Section de Mathématiques, 2-4, rue du Lièvre, C.P. 240, 1211 Genève 24, SUISSE. Peter-mann@math.unige.ch. Financed for this project by the INRIA (Oct. 2001 and Oct. 2002) and by the UHP (August 2002).

Unité de recherche INRIA Lorraine
LORIA, Technopôle de Nancy-Brabois, Campus scientifique,
615, rue du Jardin Botanique, BP 101, 54602 Villers-Lès-Nancy (France)
Téléphone : +33 3 83 59 30 00 — Télécopie : +33 3 83 27 83 19

Key-words: Error analysis, Arbitrary precision, Certified precision, Riemann zeta-function.
MSC 2000. Primary 65Gxx; Secondary 33F05, 11M06, 11Y16

Analyse d'erreur en précision arbitraire pour calculer $\zeta(s)$ avec l'algorithme de Cohen-Olivier: Description complète du cas réel et rapport préliminaire sur le cas général

Résumé : 1. Soit s un nombre réel. Si $s \geq 1/2$, $s \neq 1$ et s s'écrit en base 2 avec D_s chiffres significatifs, nous montrons que pour calculer $\zeta(s)$ avec une précision relative quelconque $P \geq 11$, c'est-à-dire pour calculer un nombre $\zeta_P(s)$ de P chiffres significatifs en base 2 de sorte que $|\zeta_P(s) - \zeta(s)|$ soit garanti inférieur au nombre $ulp(\zeta_P(s))$ representé par un "1" à la place du $P-$ème (et dernier) chiffre significatif de $|\zeta_P(s)|$, il suffit d'exécuter toutes les opérations (additions, soustractions, multiplications, divisions, calcul de k^{-s} pour des entiers $k \geq 2$) avec une précision interne

$$D = \max\left(D_s, P + \max\left(14, \left\lceil \frac{3\log P}{2\log 2} + 2.71 \right\rceil\right)\right)$$

(tout ceci contribuant une erreur inférieure à $ulp(\zeta_P(s))/2$), puis d'arrondir au nombre de P chiffres significatifs le plus proche. Par exemple si la précision finale voulue est $P = 1000$ (et si s n'a pas plus de 1018 chiffres significatifs en base 2), alors une précision interne $D = 1018$ suffit.

2. Soit $s = \sigma + it$ un nombre complexe non réel. Supposons que $\sigma \geq 1/2$ et $t > 0$. D'abord nous abordons le problème d'utiliser une erreur relative au module pour estimer les erreurs relatives à chacune des parties réelle et imaginaire du nombre calculé $\zeta(s)^*$. Connaître des régions du plan complexe où ces parties ne s'annulent pas pourrait être utile. Puis nous établissons une borne supérieure simple à calculer pour une quantité cruciale de l'analyse d'erreur (pour l'erreur relative au module), sous l'hypothèse qu'une conjecture de Brent concernant l'erreur commise lors du calcul des nombres de Bernoulli est bien vérifiée; nous remarquons que la borne supérieure que l'on obtient sans cette hypothèse peut devenir si grande pour certaine valeurs "raisonnables" de s qu'elle n'est plus d'aucune utilité pratique.

Mots-clés : Analyse d'erreur, Précision arbitraire, Précision certifiée, Fonction zêta de Riemann.
MSC 2000. Principale 65Gxx; Secondaire 33F05, 11M06, 11Y16

0.1 Preliminary note

The initial seven sections of this report (prior to which we reproduce the original abstract) will appear shortly in *Advances in Applied Mathematics* in a condensed form entitled

"On the Cohen-Olivier algorithm for computing $\zeta(s)$:
Error analysis in the real case for an arbitrary precision".$(*)$

The eighth section ("Appendix") is a preliminary report on the general case, i.e. on some of the problems we met so far for the error analysis when the argument s of $\zeta(s)$ is not a real number.

Abstract of the long version of $(*)$. Algorithms set up to compute in arbitrary precision are often not certified, in the sense that the numerical results they provide rely on lacunary error analyses, heuristic arguments, computer tests, and not on rigorous proofs. This is for instance the case for what "numerical evidence" shows is the most efficient algorithm for computing Bernoulli numbers [B], and (partially as a consequence) for a classical approximation algorithm computing the Riemann zeta-function $\zeta(s)$ [B,CO]. In the latter case, although the only published error analysis [CO] does evaluate the error of the approximation, it is not concerned by the fact that the computations required to calculate this approximation will be carried on with a finite precision arithmetic (by a computer), and thus produce other (rounding) errors.

As a first step towards clearing this matter we provide a complete error analysis of the Cohen-Olivier algorithm when the argument is real. Namely we prove that, if $s \geq 1/2$, $s \neq 1$ and s can be written with D_s bits in base 2, then in order to compute $\zeta(s)$ in any relative precision $P \geq 11$, that is, in order to compute a $P-$bit number $\zeta_P(s)$ such that $|\zeta_P(s) - \zeta(s)|$ is certified to be smaller than the number represented by a "1" at the $P-$th and last significant bit-place of $|\zeta_P(s)|$, it is sufficient to perform all the computations (i.e. additions, subtractions, multiplications, divisions, and computation of k^{-s} for integers $k \geq 2$) with an internal precision

$$D = \max\left(D_s, P + \max\left(14, \left\lceil \frac{3\log P}{2\log 2} + 2.71 \right\rceil\right)\right),$$

and then to round to the nearest $P-$bits number. For instance if the wanted precision is $P = 1000$ (and if s has no more than 1018 significant bits), then an internal precision $D = 1018$ is sufficient.

We believe that no numerical algorithm whatever should be made available before it can be based on a totally rigorous error analysis. We also believe this position is not unrealistic: indeed the algorithm we analyse in this work is now integrated in the MPFR[1]–library of the SPACES[2] project (Paris–Nancy).

[1] "Multi-Precision Floating point Reliable arithmetic".
[2] "Systèmes Polynomiaux, Arithmétique, Calculs Efficaces et Sûrs", projet INRIA.

1 Introduction

1.1 Thanks and preliminary remarks

The raw content of this work was obtained in October 2001 at the "Loria", Université Henri-Poincaré Nancy 1 (UHP), thanks to an invitation of Y.-F.S. Pétermann supported by the "Inria Lorraine", that made possible our joint contribution to its SPACES project of MPFR computer – calculation. Thanks to further invitations of the first author in 2002, supported by the UHP and the "Inria Lorraine", we completed the present work and attacked the general case where the argument s is complex (see Section 7 below). We are grateful to these institutions for their support. Our thanks to Paul Zimmermann for stimulating discussions, and for a number of suggestions and corrections he made on the successsive versions of this report. Thanks also to David Daney, who provided the error analysis for k^s (see the beginning of Section 2 below).

When we started to work on this project, we only had the limited ambition of describing one more algorithm for computing the Riemann zeta-function in arbitrary precision, and not even an original one, as we heavily rely on [CO]. But as our investigation went on, it dawned on us that in fact no complete error analysis had ever been done. It seems that in case of complex algorithms, with for instance several unbounded parameters describing the number of operations required like the p and N of (3) below in the present case, one seldom (never?) bothers to perform a precise calculation in order to determine the internal computational precision needed to certify a wanted final precision, and that one tends to (systematically does?) replace a rigorous proof by a number of computer tests based on empirical arguments. It is of course needless to say – but we shall say it anyway – that this type of procedure can provide, at best, numerical results with a very low probability of being incorrect.

1.2 The Cohen-Olivier formula

In this work we analyse the error committed while computing $\zeta(s)$ for real values of the argument s, with in addition $s \geq 1/2$. Note that for $s < 1/2$ one can appeal to the functional equation

$$\zeta(s) = 2^s \pi^{s-1} \sin\left(\frac{\pi s}{2}\right) \Gamma(1-s)\zeta(1-s) \tag{1}$$

for computing $\zeta(s)$, provided of course an algorithm for computing each of the factors of the right-hand side of (1) is available, together with a certification of the error committed.

For $s \geq 1/2$ we use the Cohen-Olivier work [CO], which is exploiting the Euler–MacLaurin summation formula applied to the real function $f(x) = 1/x^s$ for $s > 1$ and then extended by analytic continuation to the s with $|s + 2p| > 1$. This yields

$$\zeta(s) = \sum_{k=1}^{N-1} \frac{1}{k^s} + \frac{1}{2N^s} + \frac{1}{(s-1)N^{s-1}} + \sum_{k=1}^{p} \frac{B_{2k}}{2k} \binom{s+2k-2}{2k-1} \frac{1}{N^{s+2k-1}} + R_{N,p}(s), \tag{2}$$

with $|R_{N,p}(s)| < 2^{-d}$, if

$$p = \max\left(0, \left\lceil \frac{d\log 2 + 0.61 + s\log(2\pi/s)}{2} \right\rceil\right) \text{ and } N = \begin{cases} \left\lceil 2^{(d-1)/s} \right\rceil & (p = 0), \\ \left\lceil \dfrac{s+2p-1}{2\pi} \right\rceil & (p > 0), \end{cases} \tag{3}$$

where B_j denotes the $j-$th Bernoulli number (see [CO] for details).

Remark 1 *The parameter N is never very large; we have*

$$N \leq \max\left\{3, \frac{1}{2\pi}\min\{s, d\log 2 + 0.61\} + 1\right\} \quad \text{if } p = 0$$

and, for $d \geq 7$ and all values of p,

$$N \leq \frac{d\log 2 + 1.61}{2\pi} + 2.$$

Referring to (2) we compute below $\zeta_d(s) = A + B + C$ where

$$A := \sum_{k=1}^{N-1} k^{-s} + \frac{1}{2}N^{-s}, \tag{4}$$

$$B := \sum_{k=1}^{p} T_k = N^{-1-s}s\sum_{k=1}^{p} C_k\Pi_k N^{-2k+2} =: N^{-1-s}sd_{p-1}, \tag{5}$$

with

$$C_k := \frac{B_{2k}}{(2k)!}, \quad \Pi_k := \prod_{j=1}^{2k-2}(s+j) \quad \text{and} \quad T_k := N^{1-2k-s}sC_k\Pi_k,$$

and

$$C := \frac{N^{-s+1}}{s-1}. \tag{6}$$

1.3 Notation in finite precision arithmetic.

Let the (internal) computational precision be $D > d$. This means the exact numbers we work with can be written in base 2 floating-point arithmetic with (at most) D significant bits. Let now u be such a non zero real number. Then the integer $e = Exp(u)$ ("exponent" of u) and the real number $m = m(u)$ ("mantissa" of u) are uniquely defined by the equation $u = m2^e$ with $2^{e-1} \leq |u| < 2^e$ (whence $\frac{1}{2} \leq |m| < 1$). And if we define $ulp(u) := 2^{-D+e}$ ("Unit in Last Place") then we have $2^{-D}|u| < ulp(u) \leq 2^{-D+1}|u|$. In case of possible confusion with another auxiliary computational precision, we shall occasionally use the notation $ulp_D(u)$.

1.3.1 Rounding, rounding modes.

In any standard rounding mode o (towards 0, away from 0, to the left, to the right, or to the nearest), if $u = o(x)$ and $u \neq 0$ then the rounding error satisfies $|x - u| \leq ulp(u)$, and we have $|x| \leq (1 + 2^{-D+1})|u|$ and $|u| \leq (1 + 2^{-D+1})|x|$.

Similarly as for $ulp(u)$, in case of possible confusion we write $u = o_D(x)$

Although it is widely used to denote the successive roundings associated to a sequence of several operations, we chose to use the symbol $u = o(x)$ exclusively to denote one single rounding, in order to avoid confusion and mistakes. Thus we introduce the notation $u = x^*$ (or more precisely $u = x^{*_D}$), where x denotes both (1) an expression involving real (exact) numbers and operators $+, -, \times, \div$, and (2) the order in which the operations must be performed. In this expression u denotes the real number obtained after executing each operation in precision D, in the order prescribed (each real number occuring in the expression x being also rounded with precision D when used). For instance the notation $w = (x+y)^*$ means that we first compute $u = x^*$ and $v = y^*$, and then $o_D(u + v)$. Similar conventions apply to the expressions $w = (xy)^*$ and $w = (x/y)^*$.

If $u = x^*$ we also use the notation $error(u)$ (or $error_D(u)$) $= error(x^*) := |u - x|$.

1.3.2 The precision Π, and the auxiliary precisions P, d, D.

In short, our final goal is, for the value of $\zeta(s)$,
1) to certify a final precision Π in a given rounding mode.
For this, we need
2) to certify a (larger) final precision P in the rounding mode "to the nearest",
for which, in turn, we need
3) to certify a final precision $d = P + 3$ in the rounding mode "to the nearest", for the computation of each of the numbers A, B, C of (4), (5) and (6).

Finally, for this last requirement we need
4) to determine an internal computational precision D ensuring (3).

Item 1), when the rounding mode is "towards zero", is the very purpose of MPFR computer calculation. But the precision P required for that cannot be uniformly bounded in terms of Π alone. This is the "table-maker dilemma": when the old man in the clouds rounds towards zero an exact number, like $\zeta(s)$, say, he just summons its infinite base 2 expansion (beginning with the first non zero bit), and simply keeps the Π first bits; but we, poor finite creatures, have no guarantee to achieve that with any precision P "to the nearest". Of course a first try like $P = \Pi + 10$ is statistically very favourable, since unless the 9 last bits of our computed value are all equal (to 0 or 1), we are done. Otherwise we try with a larger P.

Now the choice of the computational rounding "to the nearest" is motivated by the fact that, in the error analysis below, we don't keep track of the *signs* of the errors committed while computing $\zeta_d(s)^*$. In other words whenever a rounding $x^* = o(\overline{x})$ of an $\overline{x}$ which is not representable in the internal precision is needed at some point of the algorithm, we have, with our method, no way of knowing whether $x \leq \overline{x}$ or $x \geq \overline{x}$. Since in addition $|x^* - \overline{x}|$ is likely to be, in general, much larger than $ulp_D(x^*)$, we may expect the rounding mode "to the nearest" to minimize on average the error, Hence from now on all roundings will be "to the nearest", for which we have $|x - o(x)| \leq ulp(u)/2$ and $|x| \leq (1 + 2^{-D})|o(x)|$.

We assume that the final given rounding-mode-precision we want satisfies $\Pi \geq 1$ and, as mentioned above, that our first try towards this goal is to fix $P = \Pi + 10$. If this choice turns out to be unsufficient in order to conclude, then we pick some larger P.

What we describe in this paper is the error analysis for a wanted final relative precision P, where "precision P" is here understood in the standard sense that $|\zeta(s) - \zeta_d(s)^*| < 2^{-P-1}|\zeta(s)|$ should hold. In the sequel we shall let $d = P + 3$ be the (standard) precision relative to $|\zeta(s)|$ we require for the computation of each of the expressions A, B, and C, and also such that $|R_{N,p}(s)| < 2^{-d}|\zeta(s)|$. Note that the latter is satisfied for d as in (2), since in the range considered we have $|\zeta(s)| \geq 1$ (see Lemma 0 in Section 2 below). Thus $d = P + 3 \geq \Pi + 13$ and this explains why in the sequel we assume that $d \geq 14$ and $P = d - 3 \geq 11$....And here is where the role of the symbol Π in this paper ends. To be more specific, we shall ensure below that

$$\max\left(error(A^*), error(B^*), error(C^*) + a + b\right) \leq 2^{-d}|\zeta(s)|,$$

where $a := ulp((A + C)^*)$ and $b := ulp(((A + C) + B)^*) = ulp(\zeta_d(s)^*)$.

1.4 Statement of the results

In addition to the assumptions just stated we also assume that the argument s is an *exact* number in some precision, i.e. that the expression of s in base 2 is of some *finite* length D_s. All the internal computational precisions D we use satisfy $D \geq d + 4$, $D \geq 21$, and $D \geq D_s$.

We prove the following.

Theorem 0 *Let $P = d - 3 \geq 11$. If $\zeta_d(s) = A + B + C$ as in (4), (5), and (6); if the internal precision for computing respectively A^*, B^* are respectively D_A, D_B; if the internal precision to compute C^*, and to perform the last roundings $o(A^* + C^*) = (A + C)^*$ and $o((A + C)^* + B^*) = ((A + C) + B)^* = \zeta_d(s)^*$ is D_C; if $\zeta_P(s) := o_P(\zeta_d(s)^*)$; and if*

$$D_A = \max\left(21, D_s, \Delta_A := P + \left\lceil \frac{3}{2}\frac{\log N}{\log 2} \right\rceil + 5\right),$$

$$D_B = \max\left(D_s, \Delta_B := P + 14\right) \qquad \textit{and}$$

$$D_C = \max\left(21, D_s, \Delta_C := P + \left\lceil \frac{1}{2}\frac{\log N}{\log 2} \right\rceil + 7\right);$$

then

$$|\zeta(s) - \zeta_d(s)^*| \leq 2^{-P-1.26}|\zeta(s)| \leq 2^{-P-1.25}|\zeta_d(s)^*|$$

and

$$|\zeta(s) - \zeta_P(s)| < ulp(\zeta_P(s)).$$

Note that the first conclusion ensures that the error in modulus $|\zeta(s) - \zeta_d(s)^*|$ is smaller than the number represented by a "1" at the $P+1-$st significant bit-place of the computed number $\zeta_d(s)^*$, and consequently that $|\zeta(s) - o_P(\zeta_d(s)^*)| = |\zeta(s) - \zeta_P(s)|$ is smaller than the number represented by a "1" at the $P-$th place of $\zeta_d(s)^*$. The second conclusion makes sure that $|\zeta(s) - \zeta_P(s)|$ is also smaller than the number represented by a "1" at the $P-$th (and last) place of $\zeta_P(s)$.

By using Remark 1 we can bound above the parameter N occuring in the theorem in terms of P.

Corollary 0 *If the argument s is exact in the precisions Δ_A, Δ_B and Δ_C of the theorem, then we have*

$$D_A \leq \max\left(21, P + \left\lceil \frac{3\log P}{2\log 2} + 2.71 \right\rceil\right),$$

$$D_B = P + 14, \qquad \textit{and}$$

$$D_C \leq \max\left(21, P + \left\lceil \frac{\log P}{2\log 2} + 6.24 \right\rceil\right).$$

For the implementation of the algorithm we so far simply chose the same internal precision $D := \max(D_s, \Delta_A, \Delta_B, \Delta_C)$ for the computation of A, B, and C. For values of P much larger than 1000, however, it might be worth it to use D_B for the computation of B. Note that for $P = 1000$ the corollary ensures that $D = D_A = 1018$ is adequate (provided s has no more than 1018 bits).

2 Error analysis for A

In the proof of Theorem 1 below we shall need the following estimate.

Lemma 0 *For s real, $s \geq 1/2$ we have*

$$|\zeta(s)| \geq \max\left(1, \frac{1}{1-s} - 1\right).$$

Proof. When $s > 1$ this is clear. When $1/2 \leq s < 1$ we use the inequality $\zeta(s) \leq s/(s-1) = 1 + 1/(s-1)$ (see for instance [T], Paragraph II.3.2). ◇

Note that Lemma 12 in Section 6 below yields a more precise information.

We let here the internal computational precision be $D = D_A$. Since division by 2 contributes no error, we evaluate $error(S_1^* = A^*)$, where

$$S_k := \sum_{\ell=k}^{N} \epsilon_\ell \ell^{-s} \quad (k = 1, \cdots, N).$$

and $\epsilon_\ell = 1$ ($\ell \leq N-1$), $\epsilon_N = 1/2$. We appeal to an existing algorithm for computing k^{-s} in an auxiliary internal precision D' which is slightly larger than D. If we put say $k^{-s} =: z$, z^* is here in fact $z^{*D} = o_D(z^{*D'})$, and $error(z^*) = error_D(z^*)$ denotes the error made after calculating in precision D' *and then* rounding (to the nearest in precision D). Thus $error(z^*) \leq error_{D'}(z^*_{D'}) + \frac{1}{2} ulp_D(z^*)$, and we may choose D' so as to satisfy

$$error(z^*) \leq ulp_D(z^*) \leq 2^{-D+1} z^* \leq 2^{-D+1}(1 + 2^{-D})z. \tag{7}$$

(We recall that s and $k \leq N$ are exact numbers in precision D).

[This auxiliary algorithm guarantees that

$$error((k^{-s})^{*D'}) \leq 2^{Exp((s \log N)^{*D'})+3} ulp_{D'}((k^{-s})^{*D'})$$

(personal communication from David Daney, who took care of the error analysis). On recalling that $d \geq 14$ we see with Remark 1 that $N < .28d$, whence

$$2^{Exp((s \log N)^{*D'})} \leq 2(s \log N)^{*D'} < 3s \log N < 3s \log(.28d),$$

Thus the choice

$$D' = D + 4 + \lceil (\log s + \log\log(.28d) + \log 3)/\log 2 \rceil$$

is appropriate.]

The number $D > d$ is assumed to be large enough to ensure that at every stage of the process, where we obtain $u = x^*$, say, then $error(u) \leq 2^{-d} x$. We explicitly define an adequate D at the end of this subsection. Now put

$$\gamma(d) := 1 + 2^{-d}. \tag{8}$$

Lemma 1 *Let $0 < x$, $0 < y$, $u = x^*$, $v = y^*$, $w = (x+y)^*$, and put $t_D := t2^D$. Then we have*

$$|w - (x+y)|_D \leq \gamma' x + \gamma' y + |u - x|_D + |v - y|_D,$$

where $\gamma' := (1 + 2^{-d})^2$ ($= \gamma(d)^2$).

Proof. On recalling that $(x+y)^*$ is the rounding $o(x^* + y^*)$ in precision D we have

$$\begin{aligned} ulp((x+y)^*) &= ulp(o(x^* + y^*)) \leq 2 \cdot 2^{-D} o(x^* + y^*) \\ &\leq 2 \cdot 2^{-D}(1 + 2^{-D})(x^* + y^*) \\ &\leq 2 \cdot 2^{-D}(1 + 2^{-d})^2 (x + y), \end{aligned}$$

whence

$$\begin{aligned} |(x+y) - w| &= |(x - x^*) + (y - y^*) + (x^* + y^*) - (x+y)^*| \\ &\leq |x - x^*| + |y - y^*| + |x^* + y^* - o(x^* + y^*)| \\ &\leq |x - x^*| + |y - y^*| + \tfrac{1}{2} ulp(o(x^* + y^*)) \\ &\leq (|u - x|_D + |v - y|_D + \gamma' x + \gamma' y) 2^{-D}. \end{aligned}$$

◇

Theorem 1 *For γ' as in Lemma 1 above we have*

$$error(A^*) \le \frac{3}{2}\gamma' N A 2^{-D} \le 3\gamma' N^{3/2}|\zeta(s)|2^{-D} \le (3+10^{-3})N^{3/2}|\zeta(s)|2^{-D}.$$

Proof. If we write $v_k := S_k^*$, $u_k := (k^{-s})^*$, and $w_k := ((k-1)^{-s} + S_k)^* = v_{k-1}$, then $|v_N - S_N|_D \le \gamma(d)N^{-s}/2$, and by the lemma and (7) we have

$$\begin{aligned} |v_{k-1} - S_{k-1}|_D &= |w_k - ((k-1)^{-s} + S_k)|_D \\ &\le 3\gamma'(k-1)^{-s} + \gamma' S_k + |v_k - S_k|_D, \end{aligned}$$

and

$$|v_1 - S_1|_D = |w_2 - (1 + S_2)|_D \le \gamma' + \gamma' S_2 + |v_2 - S_2|_D.$$

It is then straightforward to prove by induction that

$$error(S_1^* = A^*) = |v_1 - S_1| \le$$

$$(\gamma' + 4\gamma' 2^{-s} + 5\gamma' 3^{-s} + \cdots + (N+1)\gamma'(N-1)^{-s} + (N+1)\gamma' N^{-s}/2)\, 2^{-D},$$

where the last term inside the parentheses exists only when $N \ge 2$, the second and penultimate ones when $N \ge 3$, and the third one when $N \ge 4$. Thus $error(A^*) = 0$ if $N = 1$ and for $N \ge 2$ we have

$$error(A^*) \le \gamma'(N+1)S_1 2^{-D} \le \frac{3}{2}\gamma' N A 2^{-D} \le 3\gamma' N|\zeta(s)|,$$

where for the last estimate it is sufficient to verify that $A \le 2N^{1/2}|\zeta(s)|$. When $s > 1$ this is clear, since in fact $A < \zeta(s)$. When $1/2 \le s < 1$, we use Lemma 0. We have

$$A \le 1 + \int_1^N t^{-s} dt \le \frac{N^{1-s}}{1-s} \le N^{1-s}(|\zeta(s)| + 1) \le 2N^{1/2}|\zeta(s)|.$$

Finally since $d \ge 14$ we have $3\gamma' \le 3 + 10^{-3}$, which concludes the proof. ◇

Conclusion. Thus for computing A in a precision d relatively to $|\zeta(s)|$, it is enough to use an internal precision D_A with $(3+10^{-3})N^{3/2}2^{-D_A} \le 2^{-d}$, i.e. with $D_A - d \ge 3\log N/(2\log 2) + 1.6$. For instance

$$\text{if } \; D_A - d = \left\lceil \frac{3}{2}\frac{\log N}{\log 2} \right\rceil + 2 \;\text{ then }\; error(A^*) \le 2^{-d-0.4}|\zeta(s)|. \tag{9}$$

In the next two sections we may assume that $p \ge 1$.

3 Error analysis for C_k

The result in this part will be used in the error analysis for B. The internal computational precision is here $D = D_B$. We start with the four first coefficients

$$C_1 = \frac{1}{12}, \;\; C_2 = \frac{-1}{720}, \;\; C_3 = \frac{1}{30240}, \;\; C_4 = \frac{-1}{1209600},$$

and we shall also use $C_5 = 1/47900160$. For $k \ge 5$ the coefficients C_k are computed using the recurrence formula

$$C_k + \frac{C_{k-1}}{3!4} + \cdots + \frac{C_1}{(2k-1)!4^{k-1}} = \frac{2k}{(2k+1)!4^k}$$

(see [B], and Point 1 in Section 7 below) and the Horner type algorithm

$$C_k = \frac{-d_{k,k-1}}{24} \quad \text{with} \quad \begin{cases} d_{k,0} &= -2k, \\ d_{k,\ell} &= C_\ell + \dfrac{d_{k,\ell-1}}{4(2k-2\ell+3)(2k-2\ell+2)} \\ &=: C_\ell + e_{k,\ell} \qquad\qquad (1 \le \ell \le k-1). \end{cases} \tag{10}$$

Lemma 2 *Let $k \ge 2$. Then for each ℓ with $1 \le \ell \le k-1$ the numbers C_ℓ and $e_{k,\ell}$ are of opposite signs, C_ℓ and $d_{k,\ell}$ are of the same sign, and we have $|e_{k,\ell}| < |C_\ell|$ and $|d_{k,\ell}| < |C_\ell|$.*

Proof. The statements of the lemma clearly hold for $\ell = 1$. Now assume that the statements of the lemma hold for some $\ell \ge 1$ ($\ell \le k-2$). Then $C_{\ell+1}$ and $e_{k,\ell+1}$ are of opposite signs, since $e_{k,\ell+1}$ and $d_{k,\ell}$ are of the same sign, and since by induction hypothesis $d_{k,\ell}$ and C_ℓ are of the same sign. Now from a well-known property of the Bernoulli numbers it follows that

$$C_j = 2(-1)^{j+1}(2\pi)^{-2j}\zeta(2j) \quad (j \ge 1), \tag{11}$$

whence with (10) we have

$$\begin{aligned} |e_{k,\ell+1}| &= |d_{k,\ell}|(2k-2\ell+1)^{-1}(2k-2\ell)^{-1}/4 \\ &< |C_\ell||(2k-2\ell+1)^{-1}(2k-2\ell)^{-1}/4 \\ &= |C_{\ell+1}|\frac{\zeta(2\ell)}{\zeta(2\ell+2)}\frac{(2\pi)^2}{4(2k-2\ell+1)(2k-2\ell)} \\ &< |C_{\ell+1}|\frac{\pi^2}{6}\frac{4\pi^2}{80} < |C_{\ell+1}|. \end{aligned}$$

Finally with $d_{k,\ell+1} = C_{\ell+1} + e_{k,\ell+1}$ we see that $C_{\ell+1}$ and $d_{k,\ell+1}$ are of the same sign and that $|d_{k,\ell+1}| < |C_{\ell+1}|$. ◇

Lemma 3 *For $k \ge 5$ and $2 \le \ell \le k-1$, we have*

$$\left|\frac{e_{k,\ell}}{C_\ell}\right| \le \frac{13}{33} \quad \textit{and} \quad \left|\frac{d_{k,\ell}}{C_\ell}\right| \ge \frac{20}{33}$$

Proof. We first treat the case $\ell = k-1$; we have

$$\left|\frac{e_{k,k-1}}{C_{k-1}}\right| = \frac{||C_{k-1}| - |d_{k,k-1}||}{|C_{k-1}|} = 1 - \left|\frac{d_{k,k-1}}{C_{k-1}}\right| = 1 - 24\left|\frac{C_k}{C_{k-1}}\right| \le \frac{13}{33},$$

since by (11)

$$\left|\frac{C_k}{C_{k-1}}\right| = \frac{\zeta(2k)}{\zeta(2k-2)4\pi^2} \ge \min\left(\frac{\zeta(10)}{\zeta(8)}, \frac{1}{\zeta(10)}\right)\frac{1}{4\pi^2} = \frac{\zeta(10)}{\zeta(8)4\pi^2} = \left|\frac{C_5}{C_4}\right| = \frac{5}{198}.$$

Now if $\ell \le k-2$ we have, by Lemma 2,

$$\begin{aligned} \left|\frac{e_{k,\ell}}{C_\ell}\right| &\le \left|\frac{C_{\ell-1}}{C_\ell 4(2k-2\ell+3)(2k-2\ell+2)}\right| \le \left|\frac{C_{\ell-1}}{C_\ell}\right|\frac{1}{168} = \frac{\zeta(2\ell-2)}{\zeta(2\ell)}\frac{4\pi^2}{168} \\ &\le \zeta(2)\frac{4\pi^2}{168} = \frac{\pi^4}{252} < \frac{13}{33}. \end{aligned}$$

The other inequality now follows from $|C_\ell| = |d_{k,\ell}| + |e_{k,\ell}|$. ◇

Similarly as for A the internal (relative) precision $D = D_B$ for the computation of B is assumed to be large enough to ensure that at every stage of the process, where we obtain $u = x^*$, then $error(u) \le 2^{-.56d}x$ (where we recall that d is the target (relative) precision for the computation of each of the terms A, B, C). The choice of this exponent $-.56$ will be clear at the end of the proof of Theorem 2 below: see (13). In particular we have then

$$|u| \le \gamma|x| \quad \text{where} \quad \gamma := 1 + 2^{-.56d}. \tag{12}$$

Thus, on recalling that $d \ge 14$, we see that $\gamma \le 1.0044$. We shall verify at the end of this subsection that any $D \ge d + 4$ is appropriate.

Lemma 4 *Let $u = x^*$, $v = y^*$, assume that (12) holds for both x and y, and also that we have*

$$|u - x| \le f_x 2^{-D}|x| \quad \textit{and} \quad |v - y| \le f_y 2^{-D}|y|.$$

Then we have the following.
(1) If $w = (x + y)^*$, $|x| > |y|$, and if u and v are of opposite signs, then

$$|w - (x + y)| \le (\gamma|x| + f_x|x| + f_y|y|)\, 2^{-D};$$

(2) if $0 < x$, $0 < y$ and $w = (x/y)^*$, where y is an exact number in precision D (i.e. with at most D digits in base 2), then

$$|w - x/y| \le error(u)/y + f_x 2^{-D} x/y \le (\gamma + f_x)\, 2^{-D} x/y.$$

Proof. We first note that, by definition of "$error(o(x))$" and of "rounding to the nearest", we have $error(o(x)) \le \frac{1}{2} ulp(x) \le \frac{1}{2} ulp(o(x))$. (1) We have $|(x+y) - w| =$

$$|x - u + y - v + (u + v) - o(u + v)| \le (f_x|x| + f_y|y|)2^{-D} + |u + v - o(u + v)|.$$

The last term is

$$\le \tfrac{1}{2} ulp(u + v) \le 2^{-D}|u + v| \le 2^{-D} \max\{|u|, |v|\}$$

$$\le 2^{-D} \max\{\gamma|x|, \gamma|y|\} \le \gamma|x|2^{-D},$$

where we use assumption (12).
(2) We have $|w - x/y| \le |w - u/y| + |u - x|/y$

$$\le \frac{1}{2} ulp(\frac{u}{y}) + f_x 2^{-D}\frac{x}{y} \le 2^{-D}\frac{u}{y} + f_x 2^{-D}\frac{x}{y} \le 2^{-D}\gamma\frac{x}{y} + f_x 2^{-D}\frac{x}{y}.$$

◇

Theorem 2 *If $D \ge 21$, $D \ge d + 4$ and*

$$G_k := \begin{cases} \gamma & (1 \le k \le 4), \\ (1.67\gamma)(2.4)^k & (k \ge 5), \end{cases}$$

then we have

$$G_k \ge \frac{error(C_k^*)}{2^{-D}|C_k|} =: \overline{g}_k \quad (1 \le k \le p),$$

where p is as in (5).

Proof. When $1 \le k \le 4$ we may apply Lemma 4 (2) with $f_x = 0$, since only one exact division (by an integer with at most 13 bits) is needed. Thus we see that $G_k = \gamma \ge \overline{g}_k$ $(1 \le k \le 4)$. For $k \ge 5$, we first note that all the integers $4(2k-2\ell+3)(2k-2\ell+2)$ $(k \le p,\ 2 \le k \le p-1)$ are exact in precision D: on recalling that $D \ge 21$, $D \ge d+4$, this is easy to check with the help of (3). We consider the relative error

$$\overline{f}_{k,\ell} := \frac{error(d^*_{k,\ell})}{2^{-D}|d_{k,\ell}|},$$

and we note that $\overline{f}_{k,1} \le G_1 = \gamma$, by Lemma 4(2), with $f_x = 0$. We recursively define a sequence $\{g_k\}_{k\ge1}$ as follows:
First, $g_k := G_k = \gamma$ $(1 \le k \le 4)$. Then, for $k \ge 5$, we put $\beta = 13/20$ and

(a) $f_{k,1} = \gamma$;
(b) $f_{k,\ell} = \beta f_{k,\ell-1} + (1+\beta)g_\ell + (2\beta+1)\gamma \quad (2 \le \ell \le k-1)$;
(c) $g_k = f_{k,k-1} + \gamma$.

First a recursion argument shows that $f_{k,\ell} \ge \overline{f}_{k,\ell}$ and $g_k \ge \overline{g}_k$, and we finish the proof by checking that $G_k \ge g_k$.

As was noted before, (a) guarantees that $f_{k,1} \ge \overline{f}_{k,1}$. Then, for $2 \le \ell \le k-1$, assuming that $f_{k,m} \ge \overline{f}_{k,m}$ $(1 \le m \le \ell-1)$ we have, by using Lemma 2, Lemma 4(1) and *one* application of Lemma 4(2) (the integer $4(2k-2\ell+3)(2k-2\ell+2)$ being *exact* in precision D),

$$error(d^*_{k,\ell}) \le (\gamma|C_\ell| + (\gamma + f_{k,\ell-1})|e_{k,\ell}| + g_\ell|C_\ell|)2^{-D}.$$

Thus, with Lemma 3 and by definition (b),

$$\overline{f}_{k,\ell} \le (\gamma+g_\ell)\left|\frac{C_\ell}{d_{k,\ell}}\right| + (\gamma+f_{k,\ell-1})\left|\frac{e_{k,\ell}}{d_{k,\ell}}\right| \le (\gamma+g_\ell)\frac{33}{20} + (\gamma+f_{k,\ell-1})\frac{13}{20} = f_{k,\ell}.$$

Finally, if $\overline{f}_{k,k-1} \le f_{k,k-1}$ and if g_k is as in (c), then clearly $g_k \ge \overline{g}_k$, since C_k is obtained from $d_{k,k-1}$ with one exact division (Lemma 4(2) with $f_x = f_{k,k-1}$).

There remains to show that $g_\ell \le G_\ell = (1.67\gamma)(2.4)^\ell =: \gamma^\star\delta^\ell$. Clearly it is true if $1 \le \ell \le 4$. Suppose it is true for $1 \le \ell \le k-1$, for some $k \ge 5$. Then with (c), (b) and (a) we have

$$\begin{aligned}
g_k &= \gamma + f_{k,k-1} = \gamma + (2\beta+1)\gamma + (1+\beta)g_{k-1} + \beta f_{k,k-2} \\
&= \cdots = \gamma + (1+\beta)(g_{k-1} + \beta g_{k-2} + \cdots + \beta^{k-3}g_2) + (2\beta+1)\gamma \sum_{0\le i\le k-3} \beta^i + \gamma\beta^{k-2} \\
&\le \gamma + (1+\beta)\gamma^\star(\delta^{k-1} + \beta\delta^{k-2} + \cdots + \beta^{k-3}\delta^2) + \frac{\gamma(2\beta+1)}{1-\beta} \\
&\le \gamma\frac{2+\beta}{1-\beta} + \gamma^\star(1+\beta)\delta^{k-1}\sum_{i\ge0}\left(\frac{\beta}{\delta}\right)^i = \gamma^\star\delta^k\left(\frac{1+\beta}{\delta-\beta} + \frac{2+\beta}{1.67\delta^k(1-\beta)}\right).
\end{aligned}$$

The last factor is maximal when $k = 5$, and is then $\le .9998$.

In order to conclude the proof we still need to verify
(i) that every appeal to Lemma 4(1) we made above was legitimate, i.e. that C^*_ℓ and $e^*_{k,\ell}$ are of opposite signs for $\ell < k \le p$; and
(ii) that all the appeals to (12) we made throughout the recursive process above are legitimate for the choice of γ we made.

For (i) we show that the numbers $C^*_\ell, d^*_{k,\ell}$ and $e^*_{k,\ell}$ have the "good signs", that is those of the corresponding $C_\ell, d_{k,\ell}$ and $e_{k,\ell}$. We first note that, as $e^*_{k,\ell} =$

$o(d^*_{k,\ell-1}/(4(2k-2\ell+3)(2k-2\ell+2)))$ and $d^*_{k,\ell-1}$ must have the same sign, and as $C^*_\ell = o(-d^*_{\ell,\ell-1}/24)$ and $d^*_{\ell,\ell-1}$ must have opposite signs, it is sufficient to show that $d^*_{k,\ell} = o(C^*_\ell + e^*_{k,\ell})$ and $d_{k,\ell} = C_\ell + e_{k,\ell}$ have the same sign.

We may suppose that $d_{k,\ell} > 0$, $C_\ell > 0$, $e_{k,\ell} < 0$, the other case is treated similarly. Assuming all appeals to Lemma 4(1) made before the calculation of $d^*_{k,\ell}$ in the algorithm were legitimate, we now ensure that $d^*_{k,\ell} > 0$. As $d^*_{k,\ell} = o(C^*_\ell + e^*_{k,\ell})$ and $C^*_\ell + e^*_{k,\ell}$ must have the same sign, it is sufficient to verify that $C^*_\ell + e^*_{k,\ell} > 0$. We have

$$C^*_\ell + e^*_{k,\ell} \geq C_\ell + e_{k,\ell} - error(C^*_\ell) - error(e^*_{k,\ell}).$$

A calculation very similar to that of g_k above yields $f_{k,\ell-1} \leq \gamma^\star\delta^\ell$, that is $error(d^*_{k,\ell-1})$ $\leq \gamma^\star\delta^\ell 2^{-D}|d_{k,\ell-1}|$, whence by Lemma 4(2)

$$error(e^*_{k,\ell}) \leq \gamma 2^{-D}|e_{k,\ell}| + error(d^*_{k,\ell-1})/(4(2k-2\ell+3)(2k-2\ell+2)) \leq \gamma^\star\delta^{\ell+1}2^{-D}|e_{k,\ell}|$$

(where we used $\gamma^\star\delta^\ell + \gamma \leq \gamma^\star\delta^{\ell+1}$). Now Lemma 3 implies on the one hand that

$$error(C^*_\ell) + error(e^*_{k,\ell}) \leq \gamma^\star\delta^{\ell+1}2^{-D}(C_\ell + |e_{k,\ell}|) \leq \gamma^\star\delta^{\ell+1}2^{-D}\frac{46}{33}C_\ell,$$

and on the other hand that

$$C_\ell + e_{k,\ell} = d_{k,\ell} \geq \frac{20}{33}C_\ell.$$

Hence

$$C^*_\ell + e^*_{k,\ell} \geq \frac{20}{33}C_\ell 2^{-D}(2^D - \frac{23}{10}\gamma^\star(2.4)^{\ell+1}),$$

and since $\ell + 1 \leq p$ it is thus sufficient to check that $(23/10)\gamma^\star(2.4)^p < 2^D$. By (3) we have $p \leq (d\log 2 + .61)/2 + \pi/e + 1 \leq D\log 2/2 - \log 4 + 1.305 + \pi/e$ whence, with $\gamma < 1.0044$ and $\gamma^\star = 1.67\gamma$,

$$\frac{23}{10}\gamma^\star(2.4)^p < 10 \cdot 2^{.44D} < 2^D.$$

For (ii) it is sufficient to check that when $k \leq p$, we have $error(C^*_k) \leq 2^{-.56d}|C_k|$, provided $D \geq d + 4$. Similarly as for (i) we see that

$$1.67\gamma(2.4)^p 2^{-D} \leq 15 \cdot 2^{.44d-(d+4)} < 2^{-.56d}. \tag{13}$$

◇

4 Error analysis for B

The internal computational precision is as above $D = D_B$.

Now we compute B with the use of a Horner type algorithm. Namely $B = q_p d_{p-1}$ where d_{p-1} is the last term of the finite sequence $\{d_k\}_{k=0}^{p-1}$ defined by

$$\left\{\begin{array}{ll} d_0 & = C_p, \\ d_k & = C_{p-k} + \dfrac{d_{k-1}(s+2p-2k-1)(s+2p-2k)}{N^2} \\ & =: C_{p-k} + e_k \quad (1 \leq k \leq p-1), \end{array}\right. \tag{14}$$

where

$$q_\ell = \frac{(s+2p-2\ell-1)(s+2p-2\ell)}{N^2} \quad (1 \leq \ell \leq p-1) \quad \text{and} \quad q_p = sN^{-1-s}.$$

Note that with this notation the T_k of (6) satisfies

$$T_k = C_k \prod_{p-k+1}^{p} q_\ell. \tag{15}$$

We shall need the following estimate for $|T_k|$.

Lemma 5 *For $s \geq 1/2$ and $k \geq 1$ we have $|T_k| \leq 225e^{-2k}$.*

Proof. We have (see for instance [AS], formula 6.1.38)

$$\begin{aligned}\Gamma(s+2k-1) &\leq \sqrt{2\pi(s+2k-2)}\left(\frac{s+2k-2}{e}\right)^{s+2k-2}\exp\left(\frac{1}{12(s+2k-2)}\right)\\ &< \sqrt{2}e\left(\frac{s+2k-2}{e}\right)^{s+2k-1}\sqrt{2\pi}\frac{6}{5}\ ,\end{aligned}$$

where in the last estimate we used $e^{1/6} < 6/5$. Thus

$$\begin{aligned}|T_k| &= \left|C_k\prod_{j=0}^{2k-2}(s+j)N^{-2k-s+1}\right| = \left|\frac{2\zeta(2k)\Gamma(s+2k-1)}{(2\pi)^{2k}\Gamma(s)}N^{-2k-s+1}\right|\\ &\leq \left(\frac{4e\pi^{5/2}}{5}\right)\frac{(2\pi)^{s-1}}{\Gamma(s)}\left(\frac{s+2k-2}{2\pi eN}\right)^{s+2k-1} \leq \left(\frac{4\pi^{5/2}}{5e^{s-2}}\right)\frac{(2\pi)^{s-1}}{\Gamma(s)}e^{-2k},\end{aligned}$$

where in the last estimate we used the fact that by (3) $2\pi N \geq s+2k-1$ when $k \leq p$. In order to conclude the proof we show that for $s \geq 1/2$ the function $\phi(s) := A^{s-1}/\Gamma(s)$, where $A := 2\pi/e$, satisfies $\phi(s) \leq A^A/\Gamma(A) \leq 5.91$ for every $s \geq A-2 = .31....$ Indeed the inequality is true when $s \in [A, A+1]$; if $s \geq A+1$ then $\phi(s)/\phi(s-1) = A/(s-1) \leq 1$, whence it is also true for $s \in [A+1, A+2]$, and thus recursively for every $s \geq A+1$; and similarly if $s \leq A$ then $\phi(s)/\phi(s+1) = s/A \leq 1$, whence the inequality is true for $s \in [A-1, A]$, and in turn for $s \in [A-2, A-1]$. $\diamond$

We also need the following

Lemma 6 *Put $e_0 := 0$. Then for each $k \geq 0$ the numbers C_{p-k} and e_k are of opposite signs, C_{p-k} and d_k are of the same sign, and we have $|e_k| < |C_{p-k}|$ and $|d_k| \leq |C_{p-k}|$ ($|d_k| < |C_{p-k}|$ if $k \geq 1$).*

Proof. The statements of the lemma clearly hold for $k = 0$. Now assume that the statements of the lemma hold for some $k \geq 0$ ($k \leq p-2$). Then C_{p-k-1} and e_{k+1} are of opposite signs, since e_{k+1} and d_k are of the same sign, and since by induction hypothesis d_k and C_{p-k} are of the same sign. Now with (11) we have

$$\begin{aligned}|e_{k+1}| &= |d_k|(s+2p-2k-3)(s+2p-2k-2)/N^2\\ &< |C_{p-k}|(s+2p-2k-3)(s+2p-2k-2)/N^2\\ &< |C_{p-k-1}|\frac{\zeta(2p-2k)}{\zeta(2p-2k-2)}\frac{(s+2p-2k-3)(s+2p-2k-2)}{(2\pi N)^2}\\ &< |C_{p-k-1}|\end{aligned}$$

Finally with $d_{k+1} = C_{p-k-1} + e_{k+1}$ (and $e_{k+1} \neq 0$) we see that C_{p-k-1} and d_{k+1} are of the same sign and that $|d_{k+1}| < |C_{p-k-1}|$. $\diamond$

In the sequel we assume the hypotheses of Theorem 2, and use the following notation, where γ is as in (12).

$$\gamma_0 = (1+2^{-D}), \quad \gamma_1 := 1+2^{-D}(1+2^{-D}), \quad \gamma_2 = (1+2^{-D})^3\gamma\gamma_1^2. \tag{16}$$

We shall add below the argument s (which is assumed to be exact in precision D) to a positive integer j, also exact in precision D. The error contributed is thus only from rounding. We have

$$\begin{aligned} error(o(s+j)) &= |s+j-o(s+j)| \leq \tfrac{1}{2}ulp(o(s+j)) \leq 2^{-D}o(s+j) \\ &\leq 2^{-D}(1+2^{-D})(s+j) = \gamma_0(s+j)2^{-D}. \end{aligned} \tag{17}$$

We also need estimates very similar to those in Lemma 4 above.

Lemma 7 *Let $|x| \leq |x'|$, $u = x^*$, $v = y^*$, and assume that, for some $f_x \leq 2^D$ and $f_y \leq 2^D$ we have*

$$|u-x| \leq f_x 2^{-D}|x'| \quad \textit{and} \quad |v-y| \leq f_y 2^{-D}|y|.$$

Then we have the following.
(1) If $w = (x+y)^*$, $|x| > |y|$, if x and y are of opposite signs, and with the assumption that, in the unlikely case where $x^* = u$ and $y^* = v$ end up to be of the same sign, we set $v = 0$, then

$$|w-(x+y)| \leq error(u) + error(v) + 2^{-D}\max(|u|,|v|);$$

(2) if $w = (xy)^*$, where $y = s + j$ and s, j are exact numbers in precision D then

$$|w-xy| \leq 2^{-D}|w| + \gamma_1|u-x||y| + (\gamma_1-1)|xy|;$$

(3) if $w = (xy)^*$ then

$$|w-xy| \leq \left(\gamma_0(1+f_y2^{-D})|u/x'| + f_x + 2f_y\right)2^{-D}|x'y|;$$

(4) if $w = (x/y)^*$, where y is an exact number in precision D, then

$$|w-x/y| \leq 2^{-D}|w| + \frac{error(u)}{|y|};$$

(5) if $w = (x/y)^*$, where y is an exact number in precision D, then

$$|w-x/y| \leq (\gamma_0|u/x'| + f_x)2^{-D}|x'/y|.$$

Proof. The proof of (1) is contained in that of Lemma 4(1) above.
(2) We have $|w-xy| \leq |w-uv| + |uv-xy|$

$$\leq \frac{1}{2}ulp(w) + |v||u-x| + |x||v-y| \leq 2^{-D}|w| + \gamma_1|u-x||y| + (\gamma_1-1)|xy|,$$

where we used (17) twice (note that $\gamma_1 - 1 = \gamma_0 2^{-D}$).
(3) We have $|w-xy| \leq |w-uv| + |uv-xy|$

$$\begin{aligned} &\leq \tfrac{1}{2}ulp(w) + \tfrac{1}{2}\left\{|uv-uy| + |uy-xy| + |uv-xv| + |xv-xy|\right\} \\ &\leq 2^{-D}|w| + \tfrac{1}{2}\left\{|u||v-y| + |y||u-x| + |v||u-x| + |x||v-y|\right\} \\ &\leq 2^{-D}(1+2^{-D})|uv| + \tfrac{1}{2}\left\{|v-y|(|u|+|x|) + |u-x|(|v|+|y|)\right\} \\ &\leq 2^{-D}\gamma_0(1+f_y2^{-D})|uy| \\ &\quad + \tfrac{1}{2}\left\{f_y2^{-D}|y|(2+f_x2^{-D})|x'| + f_x2^{-D}|x'|(2+f_y2^{-D})|y|\right\}. \\ &\leq 2^{-D}|x'y|\left\{\gamma_0(1+f_y2^{-D})|u/x'| + f_x + f_y + f_xf_y2^{-D}\right\}. \end{aligned}$$

(4) and (5). We have $|w - x/y| \leq |w - u/y| + |u - x|/y$

$$\leq \frac{1}{2}ulp(w) + \frac{error(u)}{|y|} \leq 2^{-D}|w| + \frac{error(u)}{|y|}$$

and (for (5)) $|w| = o(u/y) \leq \gamma_0|u/y|$. ◇

Remark 2 *By Lemma 6 we know that $|d_k| \leq |C_{p-k}|$. In the process of estimating $error(d_k^*)$ we describe below in the proof of Theorem 3, it will be clear that the* estimate *we can obtain of $error(d_k^*)$ is* necessarily larger *than the* estimate *we use of $error(C_{p-k}^*)$ (from section 2). Therefore if, in the unlikely case we first obtain $|d_k^*| > |C_{p-k}^*|$, we then simply* replace *$|d_k^*|$ by $|C_{p-k}^*|$, there will be no risk to create a new $error(d_k^*)$ not bounded above by the* estimate *we had of the first $error(d_k^*)$. Now the proof of Theorem 3 being achieved by induction on k this shows that, with the replacement convention described above, we always have*

$$|d_k^*| \leq |C_{p-k}^*|. \tag{18}$$

So this last inequality, which is an hypothesis in the next lemma, is eventually proved to hold for every k.

Lemma 8 *Let $k \leq p-1$. Assume that the hypotheses of Theorem 2 hold, so that in particular (12) holds for $x = C_{p-k+1}$. Let γ_i $(0 \leq i \leq 2)$ be as in (16) above. Finally assume that $|d_k^*| \leq |C_{p-k}^*|$. Then we have*

(1) $$(|d_{k-1}|(s+2p-2k-1))^* \leq \gamma_2|C_{p-k+1}|(s+2p-2k-1),$$

(2) $$(|d_{k-1}|(s+2p-2k-1)(s+2p-2k))^* \leq \gamma_2|C_{p-k+1}|(s+2p-2k-1)(s+2p-2k),$$

and

(3) $$|e_k^*| \leq \gamma_2|C_{p-k+1}|(s+2p-2k-1)(s+2p-2k)/N^2.$$

We also have

(4) $$(|d_{p-1}|s)^* \leq \gamma_2|C_1|s$$

and

(5) $$(|d_{p-1}|s/N)^* \leq \gamma_2|C_1|s/N.$$

(The two last estimates will be used in the proof of Lemma 9).

Proof. (1) First we have

$$|d_k^*| \leq |C_{p-k}^*| \leq \gamma|C_{p-k}|,$$

whence by (17)

$$\begin{aligned}(|d_{k-1}|(s+2p-2k-1))^* &= o(|d_{k-1}^*|o(s+2p-2k-1)) \\ &\leq (1+2^{-D})\gamma|C_{p-k+1}|\gamma_1(s+2p-2k-1).\end{aligned}$$

(2) By using the last estimate proved we similarly obtain

$$\begin{aligned}&(|d_{k-1}|(s+2p-2k-1)(s+2p-2k))^* \\ &= o((|d_{k-1}|(s+2p-2k+1)^*)o(s+2p-2k)) \\ &\leq (1+2^{-D})^2\gamma\gamma_1^2|C_{p-k+1}|(s+2p-2k-1)(s+2p-2k).\end{aligned}$$

(3) Again with the last estimate proved, and since $o(N^2) = N^2$, we obtain (3). The proofs of (4) and (5) are very similar. ◇

Theorem 3 *Let the γ_i be as in (16) above, and let the numbers G_{p-k} ($0 \le k \le p-1$) be as in Theorem 2. Then there are numbers f_k ($0 \le k \le p-1$) satisfying*

$$f_k \ge \frac{error(d_k^*)}{2^{-D}|d_k|},$$

$f_0 = G_p$ and, for $1 \le k \le p-1$ and with $\vartheta = 2\gamma_1(2\gamma_2 + \gamma_0)$,

$$f_k|d_k| = \gamma_1^2 f_{k-1}|d_{k-1}|q_k + (\vartheta + G_{p-k})|C_{p-k}|.$$

Proof. The proof is by induction on k. Suppose that for some $k \ge 1$ there are some numbers f_j ($0 \le j < k$) satisfying the theorem, and that (18) holds for $k-1$ instead of k. Then we apply Lemma 7(2) to $x_1 = |d_{k-1}|$ and $y_1 = (s + 2p - 2k - 1)$, and then Lemma 8(1). This yields

$$\begin{aligned}
&error((|d_{k-1}|(s+2p-2k-1))^*)\\
&\le 2^{-D}(|d_{k-1}|(s+2p-2k-1))^* + \gamma_1 error(|d_{k-1}^*|)(s+2p-2k-1)\\
&\quad +(\gamma_1-1)|d_{k-1}|(s+2p-2k-1)\\
&\le (2^{-D}\gamma_2 + \gamma_1 - 1)|C_{p-k+1}|(s+2p-2k-1)\\
&\quad +\gamma_1 error(d_{k-1}^*)(s+2p-2k-1).
\end{aligned}$$

We apply again Lemma 7(2), this time to $x_2 = |d_{k-1}|(s + 2p - 2k - 1)$ and $y_2 = (s + 2p - 2k)$, and then Lemma 8(2). This yields

$$\begin{aligned}
error((x_2y_2)^*) &\le \left(2^{-D}(x_2y_2)^* + (\gamma_1 - 1)x_2y_2\right) + \gamma_1 error(x_2^*)y_2\\
&\le (2^{-D}\gamma_2 + \gamma_1 - 1)|C_{p-k+1}|(s+2p-2k-1)(s+2p-2k)\\
&\quad +\gamma_1(2^{-D}\gamma_2 + \gamma_1 - 1)|C_{p-k+1}|(s+2p-2k-1)(s+2p-2k)\\
&\quad +\gamma_1^2 error(d_{k-1}^*)(s+2p-2k-1)(s+2p-2k)\\
&\le 2\gamma_1(2^{-D}\gamma_2 + \gamma_1 - 1)|C_{p-k+1}|(s+2p-2k-1)(s+2p-2k)\\
&\quad +\gamma_1^2 error(d_{k-1}^*)(s+2p-2k-1)(s+2p-2k).
\end{aligned}$$

Now we apply Lemma 7(4) to $x_3 = |d_{k-1}|(s+2p-2k-1)(s+2p-2k)$ and $y_3 = N^2$, and then Lemma 8(3). Thus we have $x_3/y_3 = |e_k| = |d_{k-1}|q_k$, and this yields

$$\begin{aligned}
&error(e_k^*) \le 2^{-D}|e_k^*| + \frac{error(x_3^*)}{N^2} \le\\
&2^{-D}\gamma_2|C_{p-k+1}|q_k + 2\gamma_1(2^{-D}\gamma_2 + \gamma_1 - 1)|C_{p-k+1}|q_k + \gamma_1^2 error(d_{k-1}^*)q_k\\
&\le 2\gamma_1(3\gamma_2 2^{-D-1} + \gamma_1 - 1)|C_{p-k+1}|q_k + \gamma_1^2 error(d_{k-1}^*)q_k.
\end{aligned}$$

Finally we apply Lemma 7(1) to $x_4 = C_{p-k}$ and $y_4 = e_k$, and we obtain

$$\begin{aligned}
error(d_k^*) &= error((C_{p-k} + e_k)^*)\\
&\le error(C_{p-k}^*) + error(e_k^*) + 2^{-D}\max(|C_{p-k}^*|, |e_k^*|)\\
&\le 2^{-D}G_{p-k}|C_{p-k}| + 2\gamma_1(3\gamma_2 2^{-D-1} + \gamma_1 - 1)|C_{p-k+1}|q_k\\
&\quad +\gamma_1^2 error(d_{k-1}^*)q_k + 2^{-D}\gamma_2|C_{p-k}|\\
&\le 2^{-D}G_{p-k}|C_{p-k}| + 2\gamma_1(2\gamma_2 2^{-D} + \gamma_1 - 1)|C_{p-k}|\\
&\quad +\gamma_1^2 2^{-D} f_{k-1}|d_{k-1}|q_k.
\end{aligned}$$

This completes the proof of the theorem, if we note in passing that now we may assume (18) to be satisfied by k as well. ◇

Now we can estimate $error(B^*)$ in terms of f_{p-1}.

Lemma 9 *We have* $error(B^*) \le (21\gamma\gamma_0\gamma_2 + 3f_{p-1})2^{-D}B$

Proof. We recall that $B = d_{p-1}N^{-s}s/N = d_{p-1}q_p$ and use Theorem 3 and Lemmas 7 and 8. First, by Lemma 7(3) with $x = |d_{p-1}|$, $x' = |C_1|$, $y = s$, $f_x = f_{p-1}$ and $f_y = 0$, and using the fact that (18) is satisfied for $k = p-1$, we have

$$error(|d_{p-1}|s) \le (\gamma_0\gamma_2 + f_{p-1})2^{-D}|C_1|s.$$

Then, by Lemma 7(5) with $x = |d_{p-1}|s$, $x' = |C_1|s$ and $y = N$, and by Lemma 8(4), we have

$$error(|d_{p-1}|s/N) \le 2^{-D}\frac{|C_1|s}{N}(\gamma_0\gamma_2 + (\gamma_0\gamma_2 + f_{p-1})).$$

And finally by Lemma 7(3) again, with $x = |d_{p-1}|s/N$, $x' = |C_1|s/N$, $y = N^{-s}$, $f_x = 2\gamma_0\gamma_2 + f_{p-1}$ and $f_y = 2(1 + 2^{-d})$ (using (7)), and by Lemma 8(5) we have

$$\begin{aligned}\frac{error(B^*)}{2^{-D}C_1q_p} &\le \gamma_0\gamma_2(1 + 2^{-D+1}(1 + 2^{-d})) + 2\gamma_0\gamma_2 + f_{p-1} + 4(1 + 2^{-d})\\ &\le 7\gamma\gamma_0\gamma_2 + f_{p-1}.\end{aligned}$$

To complete the proof it is now sufficient to verify that $C_1q_p \le 3B$. By using Lemma 6 we see that $d_{p-1} = C_1$ if $p = 1$ and $d_{p-1} \ge C_1 + C_2(s+2)(s+1)/N^2$ if $p \ge 2$. By using (3) we thus see that $d_{p-1} \ge C_1 + C_2(s+2p-2)(s+2p-3)(2\pi)^2/(s+2p-1)^2 > 1/12 - 4\pi^2/720 > 1/12 - 1/18 = 1/36 = C_1/3$. ◇

Lemma 10 *We have*

$$B \le \frac{\pi}{6} - \frac{1}{12}.$$

Proof. First note that since $p \ge 1$, by (3) we have $N \ge (s+1)/(2\pi)$.

Now if $s \le 2\pi - 1$ then $B \le C_1|s|N^{-1-s} \le C_1(2\pi - 1) = \pi/6 - 1/12$.

And if $s \ge 2\pi - 1$ then

$$\frac{s}{N^{1+s}} \le s\left(\frac{2\pi}{s+1}\right)^{2\pi}.$$

The right-hand side of this inequality being a decreasing function of $s \ge 2\pi - 1$, it follows that

$$\frac{s}{N^{1+s}} \le (2\pi - 1)\left(\frac{2\pi}{2\pi}\right)^{2\pi},$$

whence $B \le \pi/6 - 1/12$ in this case also. ◇

Theorem 4 *If* $d \ge 14$, $D \ge 21$ *and* $D \ge d + 4$ *we have*

$$error(o^*(B)) < 1367 \cdot 2^{-D}.$$

Proof. From Theorem 3 we have, with $\alpha := \gamma_1^2$,

$$\begin{aligned}f_{p-1}|d_{p-1}| &= \alpha f_{p-2}|d_{p-2}|q_{p-1} + (\vartheta + G_1)C_1\\ &\le \alpha\Big(\alpha f_{p-3}|d_{p-3}|q_{p-2} + (\vartheta + G_2)|C_2)|\Big)q_{p-1} + (\vartheta + G_1)C_1\\ \cdots &\le \sum_1^{p-1}(G_i + \vartheta)|C_i|\alpha^{i-1}\prod_{p-i+1}^{p-1}q_\ell + \alpha^{p-1}f_0|d_0|\prod_1^{p-1}q_\ell.\end{aligned}$$

Hence, recalling (15), Lemma 5, and $f_0 = G_p$, $d_0 = C_p$, $d_{p-1}q_p = B$, we have

$$f_{p-1}B \le \sum_{i=1}^{p}(G_i + \vartheta)\alpha^{i-1}|T_i| \le \frac{225\ c'}{\alpha}\sum_{i\ge 1}(\delta\alpha e^{-2})^i + \frac{225\vartheta}{\alpha}\sum_{i\ge 1}(\alpha e^{-2})^i,$$

where $c' = 1.67c$ and $\delta = 2.4$, with $c < 1.0044$, that is

$$f_{p-1}B < 225\left(\frac{4.026}{e^2 - 2.4\alpha} + \frac{\vartheta}{e^2 - \alpha}\right).$$

Thus with Lemma 9 we have

$$error(B^*) < \left(21\gamma\gamma_0\gamma_2 B + 675\left(\frac{4.026}{e^2 - 2.4\alpha} + \frac{\vartheta}{e^2 - \alpha}\right)\right)2^{-D}.$$

If now we compute the values of the γ_i for $d = 13$ and $D = 21$, and use Lemma 10, we see that

$$error(B^*) < (9.365 + 1357)2^{-D} < 1367 \cdot 2^{-D} < 2^{-D+10.5}. \qquad \diamond$$

Conclusion. Thus, by Lemma 0, for computing B in precision d with respect to $|\zeta(s)|$ it is enough to use an internal precision D_B with $1367 \cdot 2^{-D_B} \le 2^{-d}$. For instance

$$\text{if } \ D_B - d = 11 \ \text{ then } \ error(B^*) \le 2^{-d-0.5} \le 2^{-d-0.5}|\zeta(s)|. \tag{19}$$

5 Error analysis for C, and last rounding errors

Here the internal computational precision is $D = D_C$. Note that in this section $p = 0$ is possible.

We recall that the number s is assumed to be exact in precision D (with $1/2 \le s < 2^D$). It follows that $1 - s$ and $s - 1$ are exact numbers as well in precision D. Thus by hypothesis (7) we have

$$error((N^{1-s})^*) \le ulp((N^{1-s})^*) \le \gamma(D)N^{1-s}2^{-D+1},$$

where $\gamma(D) = 1 + 2^{-D}$. Hence, by Lemma 7(4)

$$error(C^*) \le 2^{-D}|C^*| + \frac{error((N^{1-s})^*)}{|s-1|},$$

with

$$|C^*| = o\left(\frac{(N^{1-s})^*}{|s-1|}\right) \le (1 + 2^{-D})\frac{(N^{1-s})^*}{|s-1|} \le \gamma'\frac{N^{1-s}}{|s-1|},$$

where $\gamma' = \gamma(d)^2$, as in Lemma 1, whence

$$error(C^*) \le 3\gamma'\frac{N^{1-s}}{|s-1|}2^{-D}.$$

There remains to compare the respective sizes of $|\zeta(s)|$ and of $N^{1-s}/|s-1| = |C|$. For $s > 1$ we have

$$\zeta(s) = A + \frac{N^{1-s}}{s-1} + s\int_N^\infty \frac{\{t\} - 1/2}{t^{s+1}}dt,$$

(this follows for instance from formula (II.3.18)) in[T]). Thus in this case $N^{1-s}/(s-1) < \zeta(s)$. If $1/2 \le s < 1$, then by Lemma 0 $|\zeta(s)| \ge s/(1-s)$, whence $1/(1-s) \le |\zeta(s)|/s \le 2|\zeta(s)|$ and $N^{1-s}/(1-s) \le 2N^{1/2}|\zeta(s)|$. Hence we have the following.

Theorem 5 *We have*

$$error(C^*) \le 3\gamma'|C|2^{-D} \le 6\gamma'N^{1/2}|\zeta(s)|2^{-D} \le 6.1N^{1/2}|\zeta(s)|2^{-D}.$$

There finally remains to estimate the last rounding errors committed while computing $o(A^*+C^*) = (A+C)^*$, and then $o((A+C)^*+B^*) = ((A+C)+B)^* = \zeta_d(s)^*$, which we arbitrarily decide to perform *with the same internal computational precision* D_C used for computing C.

The first error is $\frac{1}{2}ulp((A+C)^*) =: a$, and the second one is $\frac{1}{2}ulp(((A+C)+B)^*) =: b$. Recalling that the number A is computed with internal precision D_A we have, for $D = D_A \geq 21$, by Theorem 1, and also using $N \leq d$ (see Remark 1),

$$A^* \leq (1 + \frac{3}{2}\gamma' N 2^{-D})A \leq 2.05 N^{1/2}|\zeta(s)|,$$

We didn't define yet the internal computational precision D_C with which C is computed, but we already know that $D_C \geq 21$, whence by Theorem 5, with $D = D_C$, we have

$$|C^*| \leq (1 + 2\gamma' 2^{-D})|C| \leq 2.05 N^{1/2}|\zeta(s)|.$$

Thus we have, with $D = D_C$

$$\begin{aligned} a &\leq 2^{-D}|o(A^* + C^*)| \leq 2^{-D}(1 + 2^{-D})(A^* + |C^*|) \\ &\leq 4.2 N^{1/2}|\zeta(s)|2^{-D}. \end{aligned} \tag{20}$$

While estimating a just above we proved that

$$|(A+C)^*| \leq 4.2 N^{1/2}|\zeta(s)|$$

(A^* being computed with internal precision D_A, and C^* and $o(A^*+C^*)$ with internal precision D_C), and by Theorem 4 and Lemma 10 we have, with $D = D_B \geq 21$,

$$B^* \leq \left(\frac{\pi}{6} - \frac{1}{12}\right) + 1367 \cdot 2^{-D} \leq .45.$$

Thus we have, again with $D = D_C$,

$$\begin{aligned} b &\leq 2^{-D}|o((A+C)^* + B^*)| \\ &\leq 2^{-D}(1 + 2^{-D})(|(A+C)^*| + B^*) \leq 4.7 N^{1/2}|\zeta(s)|2^{-D}. \end{aligned} \tag{21}$$

Finally we have, still with $D = D_C$, by Theorem 5, (20) and (21),

$$error(C^*) + a + b \leq 15 N^{1/2}|\zeta(s)|2^{-D}.$$

Conclusion. Thus, for computing C^*, and rounding then $A^* + C^*$ and finally $(A+C)^* + B^*$ in precision d with respect to $|\zeta(s)|$, it is enough to use an internal precision D_C with $15 N^{1/2} 2^{-D_C} \leq 2^{-d}$. For instance

$$\text{if } \; D_C - d = \left\lceil \frac{1}{2}\frac{\log N}{\log 2} \right\rceil + 4 \; \text{ then } \; error(C^*) + a + b \leq 2^{-d-0.2}|\zeta(s)|. \tag{22}$$

Proof of Theorem 0. Theorem 0 now follows from (2), (9), (19), and (22). Indeed since $P = d + 3 \geq 11$ we have

$$\begin{aligned} |\zeta(s) - \zeta_d(s)^*| &< 2^{-P}(2^{-3.4} + 2^{-3.5} + 2^{-3.2} + 2^{-3})|\zeta(s)| \leq 2^{-P-1.26}|\zeta(s)| \\ &\leq \frac{2^{-P-1.26}}{1 - 2^{-P-1.26}}|\zeta_d(s)^*| \leq 2^{-P-1.25}|\zeta_d(s)^*| \\ &\leq 2^{-P-1.25}(|\zeta_P(s)| + |\zeta_P(s) - \zeta_d(s)^*|) \\ &\leq 2^{-1.25} ulp(\zeta_P(s)) + 2^{-P-1.25} ulp(\zeta_P(s)) \leq (2^{-1.25} + 2^{-12.25}) ulp(\zeta_P(s)), \end{aligned}$$

whence

$$\begin{aligned} |\zeta(s) - \zeta_P(s)| &\leq |\zeta(s) - \zeta_d(s)^*| + |\zeta_P(s) - \zeta_d(s)^*| \\ &\leq (2^{-1.25} + 2^{-12.25} + 1/2) ulp(\zeta_P(s)) < ulp(\zeta_P(s)). \end{aligned}$$

6 A simpler algorithm when s is very close to 1

The analysis developed in the preceding sections is applicable for the computation of $\zeta(s)$ when s is any real number with $s \geq 1/2$, $s \neq 1$. However, when s is very close to 1 there is a much more efficient way of computing $\zeta(s)$, simply by using the approximation $1/(s-1)+\gamma$, where γ denotes in this section the Euler constant

$$\gamma := \lim_{m\to\infty}\left(\sum_{k=1}^{m} k^{-1} - \log m\right) = .577215\ldots \tag{23}$$

It is convenient to modify as little as possible the hypotheses of Theorem 0, so that for instance we assume $D \geq 21$. We have at our disposal a multi-precision–algorithm for computing γ_Δ of length Δ, for any positive integer Δ, with

$$|\gamma - \gamma_\Delta| \leq ulp(\gamma_\Delta) = 2^{-\Delta}, \tag{24}$$

and we prove the following.

Theorem 6 *Let $P = d - 3 \geq 11$ as in Theorem 0, and assume that $|s-1| < 2^{-(P+1)/2}$. Then if $D = \max(21, D_s, P+6)$, $\zeta(s)^* := o_D(o_D(1/(s-1)) + \gamma_D)$ and $\zeta_P(s) := o_P(\zeta(s)^*)$, we have*

$$|\zeta(s) - \zeta(s)^*| \leq 2^{-P-1.15}|\zeta(s)| < 2^{-P-1.14}|\zeta(s)^*|$$

and

$$|\zeta(s) - \zeta_P(s)| < ulp(\zeta_P(s)).$$

First we need the following estimate.

Lemma 11 *Let $1/2 \leq s \leq 2$. Then*

$$\left|\frac{1}{s-1}\left(\zeta(s) - \frac{1}{s-1} - \gamma\right)\right| \leq .3845$$

Proof. First assume $s = \sigma + it$ is a complex variable. Then the function

$$R(s) := \begin{cases} \dfrac{1}{s-1}\left(\zeta(s) - \dfrac{1}{s-1} - \gamma\right) & (s \neq 1) \\ \lim_{z\to 1} R(z) & (s = 1) \end{cases}$$

is analytic in the complex plane (see for instance [L, §43]). Now since $\log(m+1) - \log m \to 0$ as $m \to \infty$ we may write

$$\gamma = \sum_{k\geq 1}\left(k^{-1} - \int_k^{k+1} t^{-1}dt\right).$$

We may also write, if $\sigma > 1$,

$$\zeta(s) = \sum_{k\geq 1} k^{-s} \quad \text{and} \quad \frac{1}{s-1} = \sum_{k\geq 1}\int_k^{k+1} t^{-s}dt.$$

Hence, first for $\sigma > 1$ and then by analytic continuation for $\sigma > 0$, we have

$$R(s) \quad = \quad \frac{1}{s-1}\sum_{k\geq 1}\left(k^{-s} - k^{-1} - \int_k^{k+1}(t^{-s} - t^{-1})dt\right)$$

$$\begin{aligned} &= \frac{2^{-s}-2^{-1}}{s-1} - \int_1^3 \frac{t^{-s}-t^{-1}}{s-1}dt \\ &\quad + \frac{1}{s-1}\sum_{k\geq 3}\left(k^{-s}-k^{-1} - \int_k^{k+1}(t^{-s}-t^{-1})dt\right) \\ &=: -A(s)+C(s)-B(s) \end{aligned}$$

when $s \neq 1$, and $R(1) = -A(1)+C(1)-B(1)$ (with $X(1) = \lim_{z\to 1} X(z)$, $X = A, B, C$).

Now we restrict the argument s to the real interval $[1/2, 2]$. The function $A(s)$ is easily seen to be decreasing from $A(1/2) = \sqrt{2}-1$ to $A(2) = 1/4$, through $A(1) = \log 2/2$. In order to see that $C(s)$ also is decreasing (from $C(1/2) = 2(2(\sqrt{3}-1)-\log 3)$ to $C(2) = \log 3 - 2/3$, through $C(1) = (\log 3)^2/2)$), we write

$$C(s) = \frac{1}{s-1}\left(\log 3 - \frac{1-3^{-s+1}}{s-1}\right) = (\log 3)^2 f(-\log 3(s-1)),$$

where

$$f(u) := \frac{e^u - 1 - u}{u^2} = \sum_{n\geq 0}\frac{u^n}{(n+2)!}.$$

Now for each $s > 1$ the function $g(t) := t^{-1} - t^{-s}$, and for each $s < 1$ the function $-g(t)$, are decreasing at least for $t \geq e$, whence in particular for $t \geq 3$. It follows that in fact $B(s)$ can be written as

$$B(s) = \frac{1}{|s-1|}\sum_{k\geq 3}\left(|k^{-s}-k^{-1}| - \int_k^{k+1}(|t^{-s}-t^{-1}|)dt\right)$$

and satisfies

$$0 \leq B(s) \leq \frac{3^{-1}-3^{-s}}{s-1} =: B_1(s).$$

The function $B_1(s)$ decreases from $B_1(1/2) = 2(\sqrt{3}-1)/3$ to $B_1(1) = \log 3/3$ (the value of $B_1(2)$ is not needed). Thus we proved that

$$-.2988 < -A(1/2)+C(1)-B_1(1/2) \leq R(s) \leq -A(1)+C(1/2) < .3845$$

when $1/2 \leq s \leq 1$, and that

$$-.2809 < -A(1)+C(2)-B_1(1) \leq R(s) \leq -A(2)+C(1) < .3535$$

when $1 \leq s \leq 2$, whence the lemma. ◇

An immediate consequence of Lemma 11 is the following estimate.

Lemma 12 *If $1/2 \leq s \leq 2$ and $s \neq 1$, we have*

$$\frac{1}{s-1} \leq \zeta(s) \leq \frac{1}{s-1} + 1$$

We now complete the proof of Theorem 6. Since $|s-1| < 2^{-(P+1)/2} \leq 2^{-6} = 1/64$, Lemma 12, and Lemma 11, yield respectively

$$|\zeta(s)| > 63 \quad \text{and} \quad |\zeta(s)| > \frac{63}{64|s-1|}, \tag{25}$$

and

$$\left|\zeta(s) - \frac{1}{s-1} - \gamma\right| \leq \frac{.3845|s-1|^2}{|s-1||\zeta(s)|}|\zeta(s)| \leq .3845\frac{64}{63}2^{-P}|\zeta(s)| < \frac{8}{10}\cdot 2^{-P-1}|\zeta(s)|. \tag{26}$$

Now we have to estimate $\left|\frac{1}{s-1} + \gamma - o\left(o\left(\frac{1}{s-1}\right) + \gamma_D\right)\right| \leq$

$$\leq \left|\frac{1}{s-1} - o\left(\frac{1}{s-1}\right)\right| + |\gamma - \gamma_D| + \left|o\left(\frac{1}{s-1}\right) + \gamma_D - o\left(o\left(\frac{1}{s-1}\right) + \gamma_D\right)\right| =: a+c+b$$

With (25) we see that

$$a \leq 2^{-D}(1+2^{-D})/|s-1| \leq \frac{64}{63}2^{-D}(1+2^{-D})|\zeta(s)|,$$

and with in addition (23) that

$$b \leq 2^{-D}(1+2^{-D})(1+2^{-D}(1+2^{-D}))\left(\frac{64}{63}+.0092\right)|\zeta(s)|.$$

On recalling that $D \geq \max(21, P+6)$, and with (24), we thus have

$$\left|\frac{1}{s-1} + \gamma - o\left(o\left(\frac{1}{s-1}\right) + \gamma_D\right)\right| \leq a+b+c \leq 3.041\cdot 2^{-D}|\zeta(s)| < \frac{1}{10}\cdot 2^{-P-1}|\zeta(s)|, \tag{27}$$

The theorem now follows from (26) and (27), similarly as Theorem 0 at the end of Section 5. ◇

7 Notes on the general case

We hope to eventually return to the error analysis of $\zeta(s)$ in the general case. Things don't come out as nicely when $s = \sigma + it$ is not real, where we may suppose $\sigma \geq 1/2$ (see (1)) and $t \geq 0$ since $\zeta(\overline{s}) = \overline{\zeta(s)}$. (It should be mentioned that the values of the parameters p and N of the Cohen-Olivier formula are then much more complicated to express – see [CO]).

From some preliminary investigations it appears that, given some wanted "precision in modulus", an internal computational precision $D = D(P_0, s)$ ensuring

$$|\zeta_d(s)^* - \zeta(s)| < 2^{-P_0-1}|\zeta(s)| \tag{28}$$

might be obtained by a method inspired from that of the present paper, with the important restriction that a value of the argument s near a zero of ζ is likely to be problematic (independently from Point 1 just below). We briefly discuss the two main problems we met so far.

1. If we write $\zeta(s) = \zeta_R(s) + i\zeta_I(s)$, where $\zeta_R(s)$ and $i\zeta_I(s)$ are the real and imaginary parts of $\zeta(s)$, it seems reasonable to ask, in place of (28), for internal computational precisions Δ_R and Δ_I ensuring

$$|\zeta_R(s)^* - \zeta_R(s)| < 2^{-P-1}|\zeta_R(s)| \tag{29}$$

and

$$|\zeta_I(s)^* - \zeta_I(s)| < 2^{-P-1}|\zeta_I(s)|. \tag{30}$$

But except in special regions of the complex plane where we can establish that $|\zeta_j(s)|$ ($j = R$ or I) has a size comparable to $|\zeta(s)|$, we cannot hope to derive from (28) an explicit estimate of Δ_j. In particular, for an argument s_0 close to a zero

of $\zeta_R(s)$ or of $\zeta_I(s)$ (or of both), the requirement (29) or (30) (or both) presents a similarity with the MPFR-requirement (i.e. when a long sequence of consecutive 0s or 1s appears in the result, see Subsection 1.3): if for instance $\zeta_R(s_0)$ is small, then in order to obtain a Δ_R, the algorithm will have to increase the P_0 of (28) until $2^{-P_0-1}|\zeta(s_0)|$ is sufficiently small with respect to the $\zeta_R(s_0)^*$ of (29). In order to partially salvage our deterministic point of view it will thus be satisfactory if we can describe some large regions of the complex planes where $|\zeta_R(s)|$ does not vanish, and to provide an explicit lower bound. In this context it could for instance be worth it to provide a good estimate of $\sigma_+ := \inf\{\sigma; \sigma > 1 \text{ and } \zeta_R(s) > 0\}$. It is a triviality that $\sigma_+ < 1.8$, and with a reasonable effort one can show that $1 < \sigma_+ < 1.5$; but σ_+ is probably much closer to 1. On the other hand, the hope of describing some large regions of the complex plane where $|\zeta_I(s)|$ does not vanish appears rather compromised, since by exploiting for instance the method of proof of Theorem 8.4 in [THB] it is not difficult to verify that for every $\sigma_0 > 1$ we have $\inf_t \zeta_I(\sigma_0 + it) = 0$.

2. By analogy with the real case it is to be expected that the estimate of the sum $\sum_{k=1}^{p} \overline{g}_k |T_k|$ will be essential in the error analysis for the computation of B (see the proof of Theorem 4 above), where we recall that

$$\overline{g}_k = \frac{error(C_k^*)}{2^{-D}|C_k|} \quad \text{where} \quad C_k = \frac{B_{2k}}{(2k)!}.$$

In Theorem 2 we proved that $\overline{g}_k = O(2.4^k)$ with an explicit implied constant, but numerical evidence indicates that $\overline{g}_k$ is probably much smaller. In 1980 R.P. Brent [B] stated without proof that $\overline{g}_k = O(k^2)$, and it seems that since then this conjecture has been systematically used to compute the Bernoulli numbers in Multi-Precision packages. Which, just in passing, means that a conjectural value for the constant implied by the $O(k^2)$ must have been, somehow, determined; the issue is however not addressed in [B].

Now in the real case the term $|T_k|$ has the good taste of being extremely small, and we have at our disposal the estimate $|T_k| \le 225e^{-2k}$. This miraculously exempts us from being bothered by this problem, as our (apparently) very bad upper bound $O(2.4^k)$ for $\overline{g}_k$ is largely sufficient to ensure the convergence of the infinite series $\sum_{k\ge 1} \overline{g}_k |T_k|$.

But this miracle is unfortunately not generalizable to the complex argument $s = \sigma + it$. When $t > 0$, $|T_k|$ can become very large. One can show that

$$|T_k| \le 3.6 \cdot 2^{-d} \exp\left(2(p-k) - t \arctan\left(\frac{2(p-k)t}{(\sigma+2p-2)(\sigma+2k-2)+t^2}\right)\right),$$

and that this estimate is close to being optimal. So even the use of Brent's conjecture yields upper bounds on $\sum_{k=1}^{p} \overline{g}_k |T_k|$ already imposing a large computational precision in modulus D in some cases. And the generous use of our Theorem 2 leads to truly gigantic values of D of no practical use. Thus the help of Brent's conjecture, or possibly of a slightly weaker conjecture (with its proof as a preliminary...), appears to be unavoidable.

Appendix

8 Preliminary report on the general case

8.1 Notation and introduction

We assume that $s = \sigma + it$ with $\sigma \geq 1/2$ and $t > 0$. For $\Delta := d \log 2$ we put

$$\alpha := \Delta - .39 + \sigma \log(2\pi) - (\sigma - 1) \log |s| - \log \sigma, \quad \gamma := \frac{\alpha + \sigma}{t} - \arctan\left(\frac{\sigma}{t}\right).$$

When d a positive integer Cohen and Olivier [CO] established that (2) holds for the parameters

$$p := \begin{cases} 0 & \text{if } \sigma \geq 1 \text{ and } \gamma \leq 0, \\ 1 & \text{if } \frac{1}{2} \leq \sigma < 1 \text{ and } \gamma \leq 0, \\ \left\lceil \dfrac{1 - \sigma + t((3\gamma)^{1/3} + \gamma)}{2} \right\rceil & \text{if } 0 < \gamma < 3^{-2.5}, \\ \left\lceil \dfrac{1 - \sigma + t x_\infty}{2} \right\rceil & \text{if } \gamma \geq 3^{-2.5}, \end{cases}$$

where the real number x_∞ is computed with a low precision close to and exceeding the solution x of $x - \arctan x = \gamma$, and

$$N := \begin{cases} \left\lceil e^{\Delta/\sigma} \left(\dfrac{|s|}{2\sigma}\right)^{\frac{1}{\sigma}} \right\rceil & \text{if } p = 0, \\ \left\lceil \dfrac{|s + 2p - 1|}{2\pi} \right\rceil & \text{if } p > 0. \end{cases}$$

For $\zeta(s) = \zeta_R(s) + i\zeta_I(s)$ we shall eventually want to compute $\zeta(s)^* = \zeta_R(s)^* + i\zeta_I(s)^*$ with $\zeta_R(s)^*$ and $\zeta_I(s)^*$ in the same relative precision, P say, so that both (28) and (29), i.e.

$$error(\zeta_j(s)^*) := |\zeta_j(s) - \zeta_j(s)^*| \leq 2^{-P-1} |\zeta_j(s)^*| \quad (j = R, I)$$

hold. From which (28), i.e.

$$error(\zeta(s)^*) := |\zeta(s) - \zeta(s)^*| \leq 2^{-P_0-1} |\zeta(s)^*|,$$

easily follows for $P_0 = P$. But the converse is not true. We first intend to carry out an error analysis by considering the errors relative to modulus as in (28). This will (sometimes, see Lemma 14) be doable in a deterministic way. Then (except in very special cases, see the discussion before Lemma 15), an automatic loop in the algorithm will have to increase the value of P_0 until $2^{-P_0-1}|\zeta(s)^*|$ is sufficiently small with respect to both $|\zeta_R(s)^*|$ and $|\zeta_I(s)^*|$ to ensure that (28) implies (29) and (30). See 1 in Section 7 above.

In the two last subsections we expose with more details points 1 and 2 of Section 7, including some partial results towards the error analysis in the general case.

8.2 Remarks for the computation of A

Following Section 2 above we let the internal computational precision be $D = D_A$, and we assume that k^{-s} is computed in an internal precision $D' > D_A$ ensuring that the real and imaginary parts r and i of k^{-s} satisfy

$$error(r^*) \leq ulp_D(r^*) \leq 2^{-D+1}(1+2^{-D})r \quad \text{and} \quad error(i^*) \leq ulp_D(i^*) \leq 2^{-D+1}(1+2^{-D})i.$$

It follows that

$$error((k^{-s})^*) \leq 2^{-D+1}(1+2^{-D})|k^{-s}|.$$

As in the real case we may assume that D is large enough to ensure that at every step of the process (of computing A), at which we obtain, say, $u = x^*$, the real and imaginary parts x_r and x_i of x satisfy (8). Thus we have the following, which is proved very similarly to Lemma 1 in Section 2.

Lemma 13 *Let $x \neq 0$, $y \neq 0$, $u = x^*$, $v = y^*$, $w = (x+y)^*$, and put $t_D := t2^D$. Then we have*

$$|w - (x+y)|_D \leq \gamma'|x| + \gamma'|y| + |u-x|_D + |v-y|_D,$$

where $\gamma' := (1+2^{-d})^2$.

With this we can prove the following, very similarly to Theorem 1.

Theorem 7 *For γ' as in Lemma 13 above we have*

$$error(A^*) \leq \frac{3}{2}\gamma' N \sum_{k=1}^{N} \frac{1}{k^\sigma} 2^{-D}$$

This is of course not quite as good as Theorem 1 in which we had $error(A^*) \leq \frac{3}{2}\gamma' N A 2^{-D}$, and will not allow a deterministic treatment of the error in all cases, even if we consider errors relative to modulus. However, we have the following.

Lemma 14 *If $\sigma > 1$ then*

$$error(A^*) \leq \frac{3}{2}\gamma' N \frac{\sigma^2}{(\sigma-1)^2}|\zeta(s)|2^{-D}.$$

Proof. Since

$$\left|\frac{1}{\zeta(s)}\right| = \left|\sum_{n\geq 1} \frac{\mu(n)}{n^s}\right| \leq \sum_{n\geq 1} \frac{1}{n^\sigma} = \zeta(\sigma),$$

where μ denotes the Moebius function, and since

$$\zeta(\sigma) < 1 + \int_1^\infty t^{-\sigma} dt = \frac{\sigma}{\sigma-1},$$

we have $|\zeta(s)| > (\sigma-1)/\sigma$, whence from Theorem 7

$$\frac{2}{3}error(A^*) \leq \gamma' N \zeta(\sigma) 2^{-D} \leq \gamma' N \frac{\sigma}{\sigma-1} 2^{-D} \leq \gamma' N \frac{\sigma^2}{(\sigma-1)^2}|\zeta(s)|2^{-D} \qquad \diamond$$

Remark. We know that for $\sigma > 1 - C_1(\log t)^{-2/3}(\log\log t)^{-1/3}$ we have $|\zeta(s)| \geq C_2(\log t)^{-2/3}\ (\log\log t)^{-1/3}$ (see [THB], page 134 (Notes for Chapter 6)). If explicit values of the constants C_1 and C_2 can be obtained this could slightly increase the range of values of σ for which a deterministic treatment (for the error relative to modulus in the computation of A) is possible. Theorem 7 then yields

$$\frac{2}{3}error(A^*) \leq \frac{\gamma'(\log t)^{2/3}(\log\log t)^{1/3} N^{2-\sigma}(\epsilon_\sigma \log N + 1)}{C_2(1-\sigma)}|\zeta(s)|2^{-D}$$

with $\epsilon_\sigma = 1$ if $\sigma = 1$ and $\epsilon_\sigma = 0$ otherwise.

As has already been mentioned in Section 7 above, an explicit estimate, for the internal precision ensuring a given precision relative to the real and imaginary parts $\zeta_R(s)$ and $i\zeta_I(s)$ of $\zeta(s)$, can be obtained only in regions where $|\zeta_R(s)|$ and $|\zeta_I(s)|$

don't vanish and can be explicitly bounded below. For $|\zeta_I(s)|$ this appears hopeless, even outside the critical strip: the method of proof of Theorem 8.4 in [THB] shows that, for any $\sigma_0 > 1$ and any $\epsilon > 0$, there are indefinitely large values of t for which $|\zeta_I(\sigma_0 + it)| < \epsilon\zeta(\sigma_0)$.

It is however a triviality that

$$\sigma_+ := \inf\{\sigma; \sigma > 1 \text{ and } \zeta_R(s) = 1 + \sum_{n\geq 2} \frac{cos(t\log n)}{n^\sigma} > 0 \text{ for every } t\}$$

satisfies $\sigma_+ > 1.8$. By considering separately the minimal contributions of $\cos(t\log p)p^{-\sigma} + \cos(t\log(p^2))p^{-2\sigma}$ for p prime, $p \leq 11$, one can show that $\sigma_+ > 1.5$.

But this is tinkering and not very satisfactory. The infimum σ_+ is probably quite smaller than 1.5, but exceeds 1.

Lemma 15 *We have* $\sigma_+ > 1$.

Sketch of proof. In fact we show that $\zeta_R(s)$ has infinitely many zeros with $\sigma > 1$. Since $\log\zeta(s) = \sum_p \sum_{m\geq 1} 1/(mp^{ms})$, we have

$$\zeta(s) = \exp\left(\sum_p \sum_{m\geq 1} \frac{1}{mp^{ms}}\right) = \exp\left(\sum_p \sum_{m\geq 1} \frac{\cos(tm\log p) - i\sin(tm\log p)}{mp^{m\sigma}}\right),$$

whence

$$\zeta_R(s) = \cos\left(\sum_p \sum_{m\geq 1} \frac{\sin(tm\log p)}{mp^{m\sigma}}\right) \exp\left(\sum_p \sum_{m\geq 1} \frac{\cos(tm\log p)}{mp^{m\sigma}}\right).$$

If we put

$$f(\sigma, t) := \sum_p \sum_{m\geq 1} \frac{\sin(tm\log p)}{mp^{m\sigma}},$$

it is thus sufficient to verify that for every $\sigma_0 > 1$ close enough to 1 there is a σ' with $0 < \sigma' < \sigma_0$ and a $t_0 > 0$ such that

$$f(\sigma', t_0) - f(\sigma_0, t_0) > \pi. \tag{31}$$

This is established with Kronecker's approximation theorem. And in fact the proof establishes the existence of arbitrarily large $t_0' > 0$ satisfying (31) (for the same σ_0 and σ').

8.3 Remarks for the computation of B

As noted in Section 7, it is likely that $\sum_{k=1}^{p} \overline{g}_k |T_k|$ will be essential in the error analysis for the computation of B. But, unlike in the real case, we only obtain the following estimate (to be compared with Lemma 5)

Lemma 16 *We have*

$$|T_k| \leq 3.6 \cdot 2^{-d} \exp\left(2(p-k) - t\arctan\left(\frac{2(p-k)t}{(\sigma+2p-2)(\sigma+2k-2)+t^2}\right)\right).$$

There is no hope of any significant improvement of this estimate in general, as will be clear from the proof.

Proof. We have

$$|T_k| = \left| C_k \prod_{j=0}^{2k-2} (s+j) N^{-2k-s+1} \right| = \frac{2\zeta(2k)}{(2\pi)^{2k}} \left| \frac{\Gamma(s+2k-1)}{\Gamma(s)} N^{-2k-s+1} \right|.$$

By using

$$\left| \frac{\Gamma(x+iy)}{\Gamma(x)} \right|^2 = \prod_{n\geq 0} \left(1 + \frac{y^2}{(x+n)^2} \right)^{-1}$$

(6.1.25 of [AS]), and

$$\Gamma(x+1) = \sqrt{2\pi}\, x^{x+1/2} \exp\left(-x + \frac{\vartheta}{12x} \right) \quad (x > 0, 0 < \vartheta < 1)$$

(6.1.38 of [AS]), we obtain

$$\begin{aligned} \left| \frac{\Gamma(s+2k-1)}{\Gamma(s)} \right| &= \prod_{n=0}^{2k-2} \left(1 + \left(\frac{t}{\sigma+n} \right)^2 \right)^{\frac{1}{2}} \left| \frac{\Gamma(\sigma+2k-1)}{\Gamma(\sigma)} \right| \\ &\leq \prod_{n=0}^{2k-2} \left(1 + \left(\frac{t}{\sigma+n} \right)^2 \right)^{\frac{1}{2}} \frac{(\sigma+2k-2)^{\sigma+2k-\frac{3}{2}} e^{-2k+2+\frac{1}{6}}}{\sigma^{\sigma-\frac{1}{2}}}, \end{aligned}$$

whence, with $2\pi N \geq |s+2p-1|$ and $2\zeta(2k) \leq \pi^2/3$,

$$|T_k| \leq \left(\frac{\sigma+2k-2}{|s+2p-1|} \right)^{\sigma+2k-1} (2\pi)^{\sigma-1} \sigma^{-\sigma+\frac{1}{2}} (\sigma+2k-2)^{-\frac{1}{2}} e^{-2k+\frac{13}{6}} \frac{\pi^2}{3} \prod_{n=0}^{2k-2} \left(1 + \left(\frac{t}{\sigma+n} \right)^2 \right)^{\frac{1}{2}}.$$

Now

$$\begin{aligned} \sum_{n=0}^{2k-2} \log\left(1 + \left(\frac{t}{\sigma+n} \right)^2 \right) &\leq \int_0^{2k-2} \log\left(1 + \left(\frac{t}{\sigma+x} \right)^2 \right) dx + \log(1+(t/\sigma)^2) \\ &= -t \int_{\frac{t}{\sigma}}^{\frac{t}{\sigma+2k-2}} \frac{\log(1+w^2)}{w^2} dw + \log(1+(t/\sigma)^2) \\ &= \left(-\frac{t}{w} \log(1+w^2) + 2t \arctan w \right) \Big|_{\frac{t}{\sigma+2k-2}}^{\frac{t}{\sigma}} + \log(1+(t/\sigma)^2) \\ &= 2(1-\sigma)\log(|s|/\sigma) + 2(\sigma+2k-2)\log\left(\frac{|s+2k-2|}{\sigma+2k-2} \right) \\ &\quad + 2t \arctan\left(\frac{(2k-2)t}{\sigma^2+t^2+(2k-2)\sigma} \right). \end{aligned}$$

Hence

$$\begin{aligned} |T_k| \leq\ & \frac{(\sigma+2k-2)^{\frac{1}{2}}}{|s+2k-2|} \left(\frac{|s+2k-2|}{|s+2p-1|} \right)^{\sigma+2k-1} |s|^{1-\sigma} \sigma^{-\frac{1}{2}} (2\pi)^{\sigma-1} e^{-2k+\frac{13}{6}} \frac{\pi^2}{3} \times \\ & \times \exp\left(t \arctan\left(\frac{(2k-2)t}{\sigma^2+t^2+(2k-2)\sigma} \right) \right) \end{aligned}$$

$$:= \frac{(\sigma+2k-2)^{\frac{1}{2}}}{|s+2k-2|}\left(\frac{|s+2k-2|}{|s+2p-1|}\right)^{\sigma+2k-1} T'_k \leq \sqrt{2}T'_k$$

We first prove the lemma for $k=p$; we have

$$\sqrt{2}T'_p \leq \sqrt{2}|s|^{1-\sigma}\sigma^{-\frac{1}{2}}(2\pi)^{\sigma-1}e^{\frac{13}{6}}\frac{\pi^2}{3}\times\times\exp\left(-2p+t\arctan\left(\frac{(2p-2)t}{\sigma^2+t^2+(2p-2)\sigma}\right)\right).$$

Now as can be seen in [CO, (10)] the parameter p is chosen so as to satisfy

$$2p-1 \geq \alpha + t\arctan\left(\frac{t(2p-1)}{\sigma^2+t^2+\sigma(2p-1)}\right),$$

where we note that the argument of the arctan exceeds $t(2p-2)/(\sigma^2+t^2+\sigma(2p-2))$. Hence

$$t\arctan\left(\frac{t(2p-2)}{\sigma^2+t^2+\sigma(2p-2)}\right) \leq 2p-1-\Delta+.39-\sigma\log(2\pi)+(\sigma-1)\log|s|+\log\sigma.$$

Thus

$$T'_p \leq \sqrt{2}\sigma^{\frac{1}{2}}(2\pi)^{-1}e^{\frac{7}{6}+.39}\frac{\pi^2}{3}e^{-\Delta} \leq \sqrt{2}\frac{\pi}{6}e^{\frac{7}{6}+.39-\Delta} < 3.6e^{-\Delta},$$

and the lemma is proved for $k=p$. For other values of k we have

$$|T_k| \leq T'_k = T'_p\frac{T'_k}{T'_p} \leq 3.6e^{-\Delta}\frac{T'_k}{T'_p},$$

and

$$\begin{aligned}\frac{T'_k}{T'_p} &\leq e^{2(p-k)}\exp\left(t\left(\arctan\left(\frac{t}{\sigma+2p-2}\right)-\arctan\left(\frac{t}{\sigma+2k-2}\right)\right)\right)\\ &= \exp\left(2(p-k)+t\arctan\left(\frac{-2(p-k)t}{(\sigma+2p-2)(\sigma+2k-2)+t^2}\right)\right).\end{aligned}$$

◇

If now we assume the truth of Brent's conjecture

$$\overline{g}_k \leq ck^2, \qquad (B)$$

then from Lemma 16 we have

$$\overline{g}_k|T_k| \leq c'\exp\left(2\log(p-2(p-k)/2)+2(p-k)-t\arctan\left(\frac{2(p-k)t}{b(b-2(p-k))+t^2}\right)\right)2^{-d},$$

where we put $b := \sigma+2p-2$ and $c' := 3.6c$. Thus we may write

$$\overline{g}_k|T_k| \leq c'\exp(j(2(p-k))2^{-d},$$

where

$$j(x) = 2\log(p-x/2)+x-t\arctan\left(\frac{xt}{b(b-x)+t^2}\right)$$

The function $j(x)$ is increasing for $x<x_0$ and then decreasing, where $2t^2=(2p-2-x_0)(2p-2+\sigma-x_0)^2$. Hence we have the following.

Theorem 8 *Under the assumption of (B) and with* $x'_0 := \max(x_0,0)$*, we have*

$$\sum_{k=1}^{p-1}\overline{g}_k|T_k| \leq c'pe^{j(x'_0)}2^{-d}.$$

Example. With $\sigma = 1/2$, $t = 10'000$ and $d = 144$ we have $p = 1636$, $x_0 \simeq 2685.53$, and

$$\sum_{k=1}^{p-1} \overline{g}_k |T_k| \leq c' 2^{185-144}.$$

If we compare with the estimate we had in the real case, in the proof of Theorem 4 (where the sum corresponding to $\sum \overline{g}_k |T_k|$ converges), this means that this particular part of the error analysis will ask, in this particular case, for a contributed increase of esssentially $185 - 144 + \log c' / \log 2$ in the internal computational precision D_B to be determined. If c' is not too large this is quite reasonable. On the other hand, if instead of (B) we only use the (apparently very poor) estimate of Theorem 2 for $\overline{g}_k$, then the estimate available for the contributed increase in the precision jumps to nearly 2000!

References

[AS] M. Abramowitz and I. Stegun. *Handbook of mathematical functions.* Dover 1970 (Ninth printing).

[B] Richard P. Brent. *Unrestricted algorithms for elementary and special functions.* Information Processing 80 (ed. S.H. Lavington), North-Holland, Amsterdam 1980. Retyped with minor corrections in 1999.

[CO] Henri Cohen et Michel Olivier. *Calcul des valeurs de la fonction zêta de Riemann en multiprécision.* C.R. Acad. Sci. Paris **314** (1992), 427-430.

[L] E. Landau. *Handbuch der Lehre von der Verteilung der Primzahlen.* Leipzig und Berlin, Teubner 1909; third edition, Chelsea 1974.

[T] Gérald Tenenbaum. *Introduction à la théorie analytique et probabiliste des nombres.* Cours spécialisés. Collection de la Société Mathématique de France No 1. 1995.

[THB] E.C. Titchmarsh. *The theory of the Riemann zeta-function.* Oxford, Clarendon Press 1951; second edition revised by D.R. Heath-Brown, ibid. 1986.

Unité de recherche INRIA Lorraine
LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399